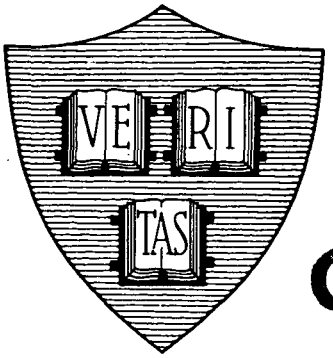
SOME SEQUENTIAL, DISTRIBUTION-FREE PATTERN CLASSIFICATION PROCEDURES WITH APPLICATIONS

Technical Report No. 2



By

J. L. Poage

July 1971

NATIONAL AERONAUTICS AND SPACE ADMINISTRATION

Prepared under Grant NGL 22-007-143

Division of Engineering and Applied Physics

Harvard University • Cambridge, Massachusetts

SOME SEQUENTIAL, DISTRIBUTION-FREE PATTERN CLASSIFICATION PROCEDURES WITH APPLICATIONS

By

J. L. Poage

Technical Report No. 2

July 1971

Prepared under Grant NGL 22-007-143
Division of Engineering and Applied Physics
Harvard University Cambridge, Massachusetts

for

NATIONAL AERONAUTICS AND SPACE ADMINISTRATION

SOME SEQUENTIAL, DISTRIBUTION-FREE PATTERN CLASSIFICATION PROCEDURES WITH APPLICATIONS

By

J. L. Poage

Division of Engineering and Applied Physics
Harvard University Cambridge, Massachusetts

ABSTRACT

Some sequential, distribution-free pattern classification techniques are presented. In many classification problems, the observations on which the classification decision is to be based are costly to measure. A sequential test seems appropriate since observations are measured only until enough information is known to make a decision with a certain level of confidence. Also in many cases, the only information available about the pattern classes is a set of training samples from each class. Since the underlying probability density functions are unknown, distribution-free classification methods are needed. The specific decision problem to which the proposed classification methods are applied is that of discriminating between two kinds of electroencephalogram (EEG) responses recorded from a human subject - spontaneous EEG and EEG driven by a stroboscopic light stimulus at the alpha frequency. Sequential, distribution-free methods are suitable since it is generally desired to terminate the EEG recording

as quickly as possible and since there is no knowledge of probability density functions underlying the EEG waveforms.

The classification procedures proposed make use of the theory of order statistics. Estimates of the probabilities of misclassification are given. One of the methods presented is an estimated version of the Wald sequential probability ratio test (SPRT). This method utilizes density function estimates, and in formulating this test, a new probability density function estimate is proposed. Convergence in probability of the estimate to the true density function is shown. The other method presented is a sequential version of the separating hyperplane approach to pattern classification.

The procedures were tested on Gaussian samples and on the EEG responses. Smaller error rates were easier to obtain with the estimated SPRT. In particular, error rates as low as .1% were obtained. With sequential tests, it is possible to specify the probability of error decisions before the test is conducted, and the experimental error rates of the procedures agree with the specified error probabilities.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER I

INTRODUCTION

I.1 Pattern Classification Problem

In the pattern classification problem, a pattern is given that was drawn from one of several pattern classes, and a decision must be made as to which class the pattern was drawn. In order to classify the pattern, a way must be found to characterize the pattern, and then a method must be developed of processing the characterization of the pattern to classify it. It is usual to attempt to characterize the pattern as a set of s real numbers $x = (x^1, x^2, \ldots, x^s)$. The components x^i of the pattern vector are called features and are usually measurements of various attributes of the pattern. The choice of features to characterize the pattern is called the feature extraction problem. While any number of pattern classes is possible, this report will consider only classification problems with two pattern classes C^1 and C^2. Once the observation x has been characterized as a vector, the problem of classifying x can be formulated as finding a scalar function g(x) such that x is classified as coming from C^1 if $g(x) < 0$ and as coming from C^2 if $g(x) > 0$.

In viewing the classification problem geometrically, each pattern has been considered as a point in an s-dimensional space. Thus $g(x) = 0$ is a separating surface that divides the sample space into two regions corresponding to classifying the pattern x as coming from C^1 or C^2.

In most meaningful classification problems, the two pattern classes overlap to some extent and so are not separable in the s-dimensional space. The objective in this case is to construct a classification procedure that is optimal in some sense as regards misclassifications.

Since a pattern can be treated as a set of real numbers, the two pattern classes will be characterized in this report by the probability density functions $f(x|C^1)$ and $f(x|C^2)$. This does not mean the density functions are always known but means that the patterns from each class can be treated as random variables with a particular probability density function. It may be that the density functions reflect noise in measuring the features, or it may be that the patterns themselves follow a particular density function.

Before proceeding to a more detailed discussion of pattern classification methods, an example of a classification problem will be given.

I.2 Electroencephalograms

The application of pattern classification techniques to the biomedical field has received increasing attention in recent years. One specific area that has been studied is that of making decisions about the state of a patient based on electroencephalograms (EEG). An EEG is a recording of the electrical activity of the brain. From the EEG waveform, some assessment can be made on the state of the patient; for example the level of consciousness of the patient can be determined

or some pathelogical conditions of the brain can be detected. The electrical activity is measured by electrodes on the surface of the scalp, and the EEG wave is generally considered to be a recording of the gross activity of a large number of cells. An EEG response can thus be considered to be a sample from a random process. The pattern classification aspect of the problem now becomes apparent.

An EEG measured from a patient placed in a darkened, soundless room isolated from external stimuli is called a spontaneous EEG. If a light is flashed periodically into the patient's eyes, the resulting EEG wave between two consecutive flashes is called an evoked response. This report will treat a classification problem to determine whether given EEG responses are spontaneous or evoked.* As mentioned previously, in order to classify an EEG wave, a set of features to describe the wave must be extracted, and a decision rule to classify the set of features must be formulated.

I.3 Feature Extraction

Prabhu [1] has written a paper that discusses feature extraction for the EEG classification problem. As recorded from the patient,

---

*Although the flashing of the light can be readily detected by merely observing the light, this thesis attempts to make the decision on the light by observing an EEG response from the patient. The decision problem considered here is a first step toward more meaningful problems such as determining the level of unconsciousness of a patient during surgery. An unconscious patient would react differently to a light stimulus than an awake patient.

the EEG is a continuous waveform of the amplitude of the electrical activity. The response between two consecutive flashes of the light is considered to be one sample. To facilitate the use of a digital computer, the amplitude was sampled in time at a set frequency so that each sample EEG response was a vector. If the sampling rate is high, the dimension of the sample vector may be quite large. Since the complexity involved in finding a suitable decision rule increases as the dimension of the sample increases, a subset of the features may be selected to be used in the decision rule. Prabhu [1] has developed a feature reduction scheme that picks a subset of the total number of features. The features in the subset are selected according to their effectiveness in some sense for classification purposes. This feature reduction method is discussed in detail in Appendix II.1 and in Prabhu [1].

I.4 Structure of the Classification Problem

Now that a set of features has been extracted so that the EEG responses can be represented as vector samples, a decision process for classifying the EEG samples must be developed. The purpose of this report is to develop some classification techniques that are applicable to a class of problems represented by the EEG decision problem. Before discussing the specific properties of this class of problems, some general considerations of classification problems will be presented.

In classifying an observation x, the two types of errors possible are to decide $x \in C^2$ when actually $x \in C^1$, called error of type I,

and to decide $x \in C^1$ when actually $x \in C^2$, called error of type II. Criterions for evaluating the effectiveness of decision rules are usually expressed in terms of the probabilities of these error occurring. Let α = p(error of type I) and β = p(error of type II). Three examples of criterions expressed in terms of α and β follow.

1.) If the prior probabilities of an observation coming from C^1 or C^2, $p(C^1)$ and $p(C^2)$ respectively, are known, then an expected loss function associated with a misclassification can be expressed as

$$E(\text{loss of misclassification}) = L_1 \alpha p(C^1) + L_2 \beta p(C^2)$$

where L_1 and L_2 are the cost of errors of type I and type II. A possible criterion is to formulate a decision rule to minimize the expected loss function. The Bayes test [2] satisfies this criterion.

2.) Another possible criterion is to require that α be below a specified value and then minimize β. This criterion is followed by the Neymann-Pearson test [3].

3.) If the number of observations drawn before making a decision is variable and not predetermined, a decision rule can be devised where both α and β are below specified values. The Wald sequential probability ratio test [4] satisfies these conditions and minimizes the expected number of observations needed for a decision.

Another factor that influences the choice of methods for solving a classification problem is the type of information known about the

two pattern classes. When the probability density functions describing the pattern classes are known, there are many well-known decision tests that can be used, such as those already mentioned. In many cases, however, the density functions are unknown, but sets of samples drawn from each class are known. These sets of samples from each class are called training sets. When training sets are the only information available, pattern classification techniques must be formulated from the training sets without using the density functions.

The development of a decision procedure then depends on two factors:

1.) the information known about the two pattern classes, and

2.) the criterion.

The choice of a criterion is influenced by the information available, e.g. if the density functions are unknown it is not possible to minimize the actual probability of a misclassification but only perhaps an estimate of it. The criterion also embodies the characteristics that are important to a particular decision problem, such as the number of observations that may be taken before a classification decision is made.

I.5 The Approach Taken in this Report

In the classification problem of the EEG waves mentioned in Section I.2, the underlying density functions of the EEG waves are unknown. But it is generally possible to record a series of EEG responses from the patient to use as training sets. Pattern classification procedures

that do not involve knowledge of the underlying density functions are called distribution-free. The techniques proposed in this report are distribution-free.

In making medical tests on a patient, the measurements are often costly and discomforting to the patient. Thus it seems desirable to terminate the measurements as quickly as possible, but at the same time the final decision on the state of the patient must be made with a certain level of confidence. A sequential test appears appropriate for many bio-medical classification problems since observations are taken one at a time only until enough information is known to make a decision with a certain level of confidence. In sequential tests, the p(error of type I) and p(error of type II) can both be specified before the test. Sequential tests are suitable for the EEG decision problem since the stroboscopic light can be flashed and responses sampled on demand until enough data has been gathered to make a decision.

As mentioned in the previous paragraph, sequential methods take observations one at a time until the string of observations provides enough information in some sense to classify the observations. If the observations are vectors, a whole new vector observation of the several features is taken. After each observation is taken, three outcomes are possible:

1.) decide the observations taken so far are from C^1

2.) decide the observations taken so far are from C^2

3.) decide to take another observation since not enough information is known to make a decision.

Stated analytically, the classification problem using the sequential method is to find a scalar function and two thresholds such that after

t observations have been taken

$$g(x_1,x_2,\ldots,x_t) \leq B \qquad \text{decide } C^1$$

$$B < g(x_1,x_2,\ldots,x_t) < A \qquad \text{take another observation}$$

$$g(x_1,x_2,\ldots,x_t) \geq A \qquad \text{decide } C^2 \ .$$

Since the two thresholds can be set independently, it is possible to construct a sequential test where the p(error of type I) and p(error of type II) are both specified to be certain values. As an example, consider Figures I.1 and I.2. In the test using one observation shown in Figure I.1, two outcomes are possible, and a decision is made according to which side of a single threshold the observation lies. Since only one threshold is used, the probabilities of type I and type II errors cannot be set independently. In the sequential method of Figure I.2, three outcomes are possible after each observation is taken. The two thresholds that separate the three decision regions can be set independently, and hence the probabilities of errors of type I and type II can both be set to specified values. Since only enough observations are taken to make a decision with the confidence that the p(error of type I) and p(error of type II) have certain values, the sequential method has the merit that test procedures can be constructed which require, on the average, fewer observations than equally reliable test procedures based on a predetermined number of observations [4].

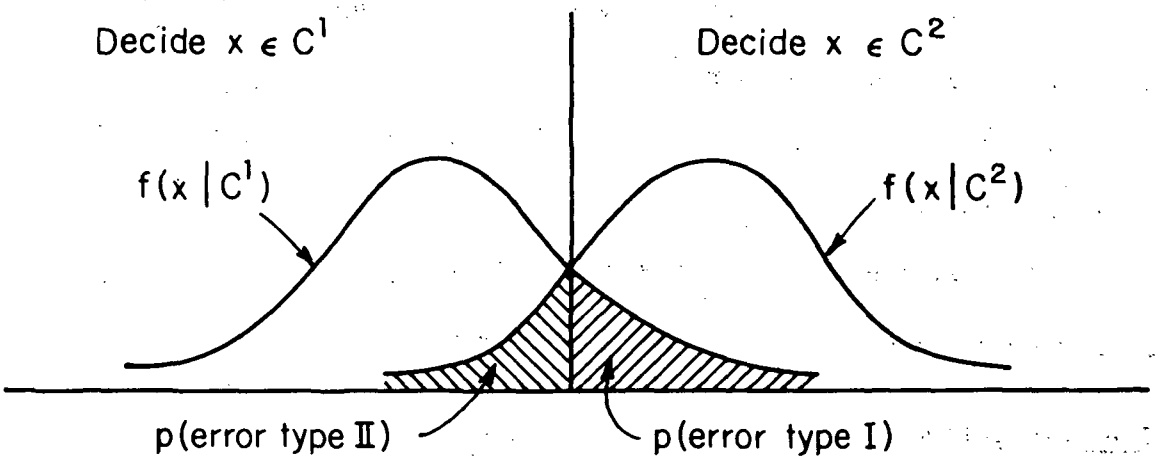


FIGURE I.1

Error Probabilities for Testing One Observation

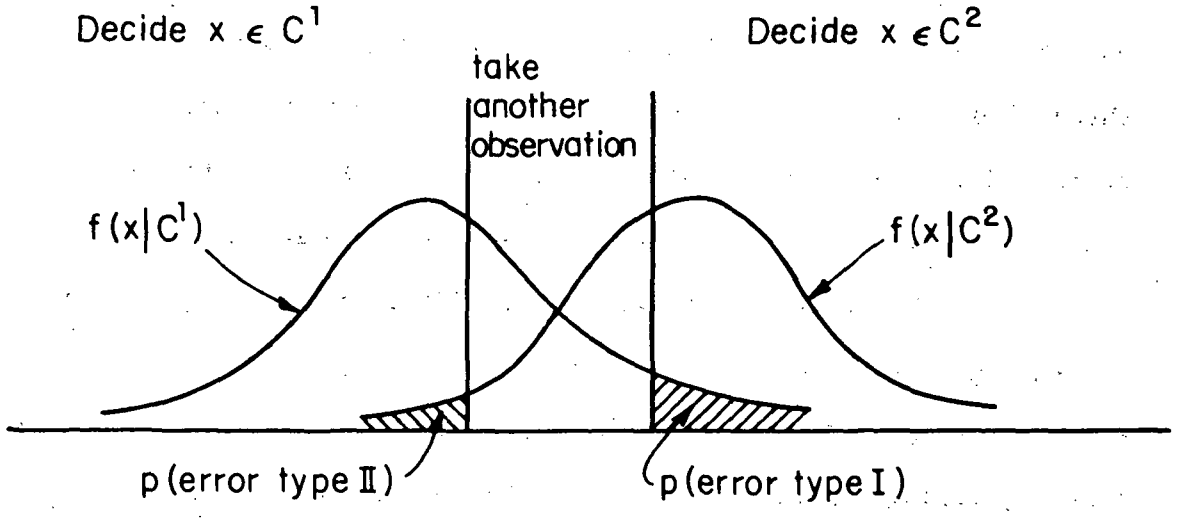


FIGURE I.2

Error Probabilities for Sequential Test

The classification procedures proposed in this report are distribution-free and sequential. The methods are applicable in classification problems where:

1.) the density functions of each class are unknown but training sets are known, and

2.) a string of a variable number of observations, all from the same unknown class, can be sampled on demand.

I.6 General Outline

Two types of sequential, distribution-free procedures are presented in the chapters that follow. In one, a series of thresholds are calculated from training samples, and each observation that is taken in the sequential sampling is compared to a different pair of thresholds depending on the number of the iteration. In the other approach, the same pair of thresholds is used throughout the sequential procedure, and the scalar function of the observations that is compared to the thresholds is altered at each iteration to include the information contained in the new observation. Chapter II describes the former approach, and Chapters III through VI are concerned with the latter.

Chapter II presents a brief review of the theory of order statistics and then uses some results from order statistic theory to calculate a set of thresholds for a sequential test. The thresholds are calculated from the training sets in such a way that an estimate of the probability of a misclassification is obtained. Multidimensional samples are treated by transforming them into scalars with a linear transformation.

Experimental results are shown for the procedure tested on both Gaussian and EEG data.

Chapters III through VI are concerned with the estimation of probability density functions from training samples and the use of density estimates in a sequential test called the sequential probability ratio test (SPRT). The SPRT utilizes the ratio of the two density functions representing the pattern classes. The ratio of the densities is evaluated at the values of the observations and compared to two thresholds. Since the density functions are unknown in the problems considered in this thesis, estimates of the densities are used in the SPRT. Chapter III discusses some approaches for estimating density functions and surveys several known estimates. A new density estimate is proposed in Chapter IV. The estimate is of a step-function form where the boundaries of the steps are determined by the training samples. The estimate is shown to converge in probability to the true density as the number of training samples tends to infinity.

Chapter V begins with a discussion of the SPRT, and then formulates an estimated version of the SPRT with the new density estimate. The new density estimate was chosen because of its low computer storage requirement and ease of calculation. Experimental results are shown for independent Gaussian samples. Some techniques for handling multi-dimensional samples and dependent observations are discussed in Chapter VI. The methods involve taking a linear combination of the features of multi-dimensional samples or taking the sum of several dependent observations so that only scalar samples are considered. The procedures are tested on EEG data.

CHAPTER II

A SEQUENTIAL DISTRIBUTION-FREE PATTERN CLASSIFICATION PROCEDURE USING ORDER STATISTICS

This chapter presents a sequential, distribution free pattern classification procedure that makes use of some results from order statistics. The material in this chapter is self-contained, and future chapters do not depend on what is developed here.

II.1 Introduction

The algorithm that follows assumes the type of prior information and criterion listed in Section I.5 namely that a training set from each class is known and the test is to be sequential. One popular method of solving the classification problem with training sets is to place a hyperplane between the two sets of training samples that separates the two classes of samples as much as possible. An observation is classified according to which side of the hyperplane it lies. Generally such algorithms provide no direct estimate of the probability of misclassification, and the decision is made based on examining only one observation. Henrichon and Fu [5] have formulated an algorithm which partitions the sample space into decision regions by training on sample sets of known classification and uses order statistics to find an upper bound on the misclassification probability. This chapter presents a method which attempts to improve the error in classifying observations from inseparable classes by taking several observations before deciding on classification. The observations are drawn sequentially. A distribution-free estimate of the probability of misclassification is presented. The remainder of the chapter describes the algorithm and experimental results.

II.2 Assumptions

The method is designed to decide if an unknown observation belongs to one of two classes which shall be referred to as class 1 and class 2. The algorithm is trained on sample sets of known classification and is distribution free. The following assumptions are made about the samples:

i. that a training set from each class is known

ii. that the samples are independently, identically distributed in each class

iii. that the random variables from each class are of the continuous type*

(thus the probability of any two samples being equal is zero)

iv. that several observations, all from the same unknown class to be classified, can be taken since the method is to be sequential.

II.3 Order Statistics and Ordering Functions

Several properties of order statistics are used in this chapter. A brief presentation of order statistics, including some distribution-free properties, is given in this section without proof. Appendix II.3 may be consulted for a more detailed discussion of order statistics.

---

*A random variable is of the continuous type if the distribution function $F(x)$ is everywhere continuous and the density function $f(x) = F'(x)$ exists and is continuous for all x, except possibly at certain points of which any finite interval contains at most a finite number. Thus $F(x) = p(\eta \leq x) = \int_{-\infty}^{x} f(t)dt$ [6]. A function $F(x)$ which has these properties is said to be absolutely continuous.

Let $X_1 X_2, \ldots, X_n$* be a set of n independent scalar random variables from a continuous probability distribution function F(x). The samples can be arranged in ascending order, $X_{i_1} < X_{i_2} < \ldots < X_{i_n}$. For convenience, let the samples be relabeled, $Y_1 = X_{i_2}, Y_2 = X_{i_2}, \ldots, Y_n = Z_{i_n}$, so that $Y_1 < Y_2 < \ldots < Y_n$. In the set $(Y_1, Y_2, \ldots, Y_n)$, each member Y_i is called an order statistic. If X is a scalar random variable, F(X) is also a random variable. The random variable F(X) turns out to have a uniform distribution on the interval (0,1). Recall that the random variable F(X) can take on values between 0 and 1, and $F(X) = p(\eta \leq X)$. So it is equally likely for any random sample X that $p(\eta \leq X)$ be anywhere between 0 and 1. The expectation of $F(Y_j) - F(Y_i)$ can be shown to be

$$E[F(Y_j) - F(Y_i)] = \frac{j-i}{n+1} \qquad j > i \quad . \tag{II.1}$$

Thus

$$E[F(Y_{j+1}) - F(Y_j)] = \frac{1}{n+1} \quad . \tag{II.2}$$

It is observed that n random variables thus arranged in ascending order partition the density function into n+1 parts. The expected value of the probability of a sample falling between any two neighboring order statistics is 1/(n+1). The variance of $[F(Y_j)-F(Y_i)]$ can be shown to be

$$E[(F(Y_j)-F(Y_i))-E(F(Y_j)-F(Y_i))]^2 = \frac{(j-i)(n-j+i+1)}{(n+1)^2(n+2)} \quad . \tag{II.3}$$

For dealing with multi dimensional samples, ordering functions are used to transform the vector samples onto the real line. Let X be a multidimensional random variable with a continuous distribution

---

*Random variables are denoted by capital letters.

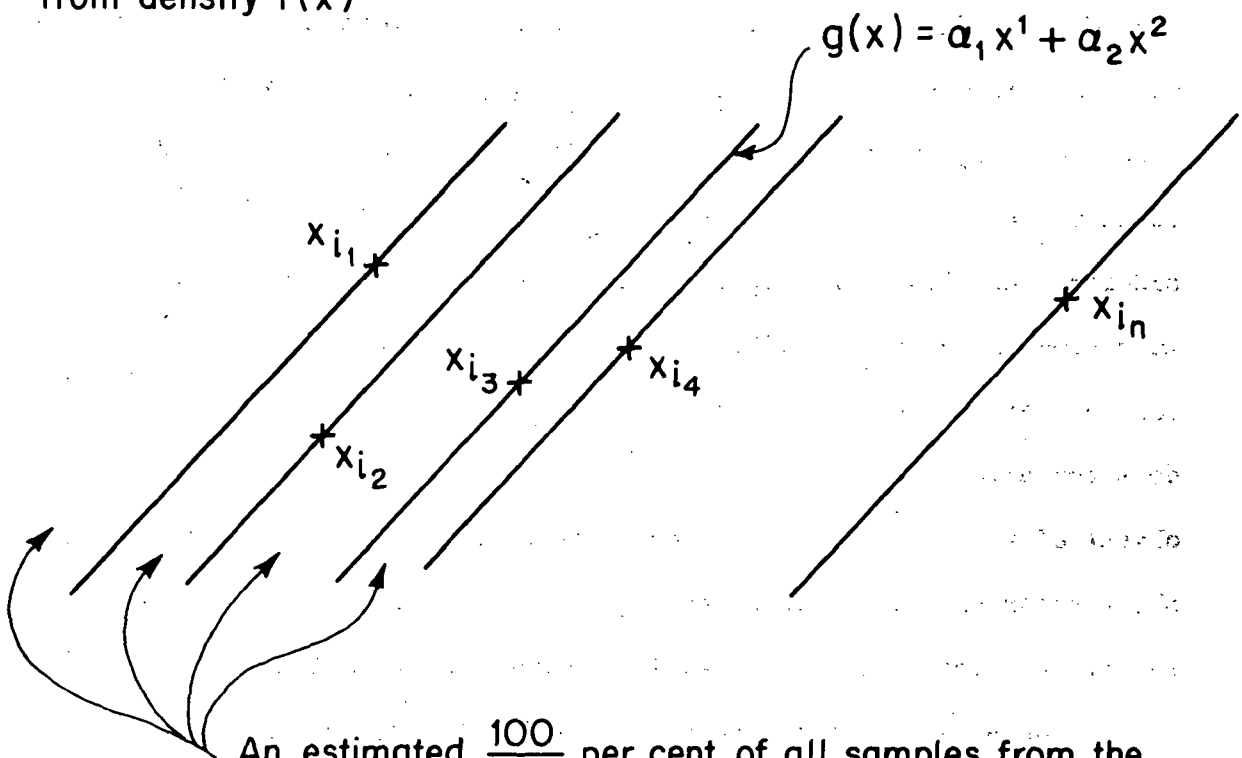
function F(x). If W = g(X) is a random variable with a continuous distribution function G(w), then g(x) is an ordering function. Kemperman [7] has shown how the sample space can be partitioned using a class of ordering functions so that the distribution of the probability of a future observation falling in any partition can be found. An example of using one linear ordering function for partitioning the sample space is given in Figure II.1. For the random sample $X_1, X_2, \ldots, X_n$ from the multivariate, absolutely continuous distribution function F(x), if the transformed vectors are ordered, $g(X_{i_1}) < g(X_{i_2}) < \ldots < g(X_{i_n})$ and relabeled, $W_1 = g(X_{i_1})$, $W_2 = g(X_{i_2}), \ldots, W_n = g(X_{i_n})$ then

$$E[G(W_j)-G(W_k)] = \frac{j-k}{n+1} \qquad j > k \tag{II.4}$$

$$= E[p(g(x_{i_k}) < g(x) < g(x_{i_j}))]$$

The expected probability of a future observation falling in the block partitioned by $g(x_{i_j})$ and $g(x_{i_k})$ is $\frac{j-k}{n+1}$ for $j > k$. For example, let $g(x^1, x^2, \ldots, x^s) = \alpha_1 x^1 + \alpha_2 x^2 + \ldots + {}_s x^s$ be a linear function and let $x_1, x_2, \ldots, x_n$ be a set of n vector samples. Then if the transformed samples are arranged so that $g(x_{i_1}) < g(x_{i_2}) < \ldots < g(x_{i_n})$, then the expected probability of a future observation falling in the segment between the planes $g(x_{i_j})$ and $g(x_{i_{j+1}})$ is $\frac{1}{n+1}$ independent of the choice of g as well as the underlying distribution for x. Ordering functions and order statistics are discussed more fully in Appendix II.3.

Given sample set $x_1, x_2, \cdots, x_n$ of two dimensional vectors from density $f(x)$



An estimated $\frac{100}{n+1}$ per cent of all samples from the density $f(x)$ lie within each segment.

Figure II.1

Example of Linear Ordering Function

II.4 The Algorithm

II.4.1 Use of Two Thresholds

In dealing with multidimensional samples, this chapter uses the same ordering function throughout for any one testing procedure. The use of a single ordering function may not be optimal for many data sets, but for some unimodal densities with one region of overlap the shapes of the data sets are such that the use of a linear ordering function sufficiently separates the two classes. Utilizing different ordering functions for different iterations requires considerably more computation and is discussed further in Section II.7. Of course, for scalar samples the question of an ordering function does not arise. For whatever ordering function is chosen, the object of the algorithm is to decide to which class an unknown observation belongs so the ordering function chosen should separate the two classes of training samples as much as possible.

A convenient type of ordering function to use is a linear function. The distribution of the linearly transformed samples is continuous. Figures II.2 and II.3 show two examples of linear ordering functions. The two training sets in the figures cannot be separated by a linear function. The function $g_2(x)$ of Figure II.3 separates to a greater degree the two classes of training samples than the function $g_1(x)$ of Figure II.2. For a decision algorithm, the ordering function $g_2(x)$ is the better choice.

Many algorithms exist which yield a single linear separating plane between the two classes of training samples. Ho and Agrawala [2]

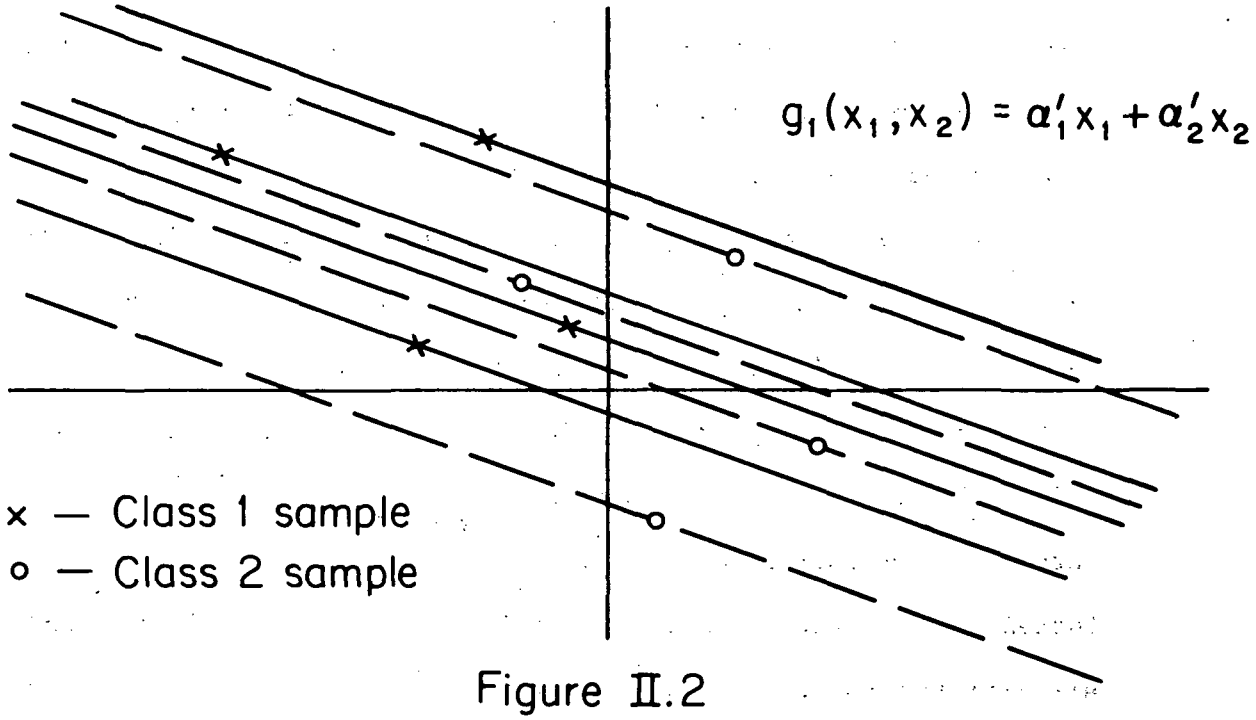


Figure II.2

Linear Ordering Function with Poor Separating Qualities

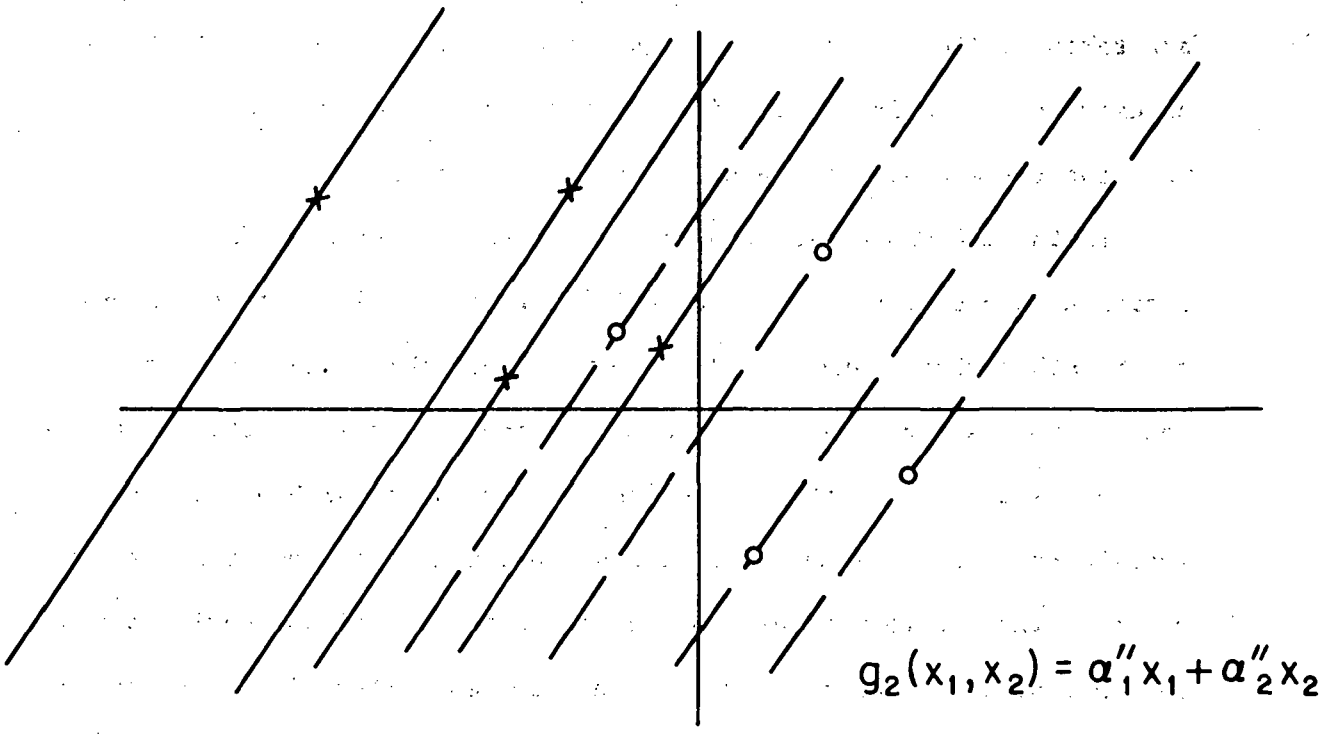


Figure II.3

Linear Ordering Function with Good Separating Qualities

give a survey of many linear separating algorithms. The equation of such a separating hyperplane can be used as an ordering function since it has good separating qualities.

When a single ordering function is used on all training samples the expected probability of a new sample falling in the segment between any two planes, each placed through a training sample, is the same as the expected probability of the transformed sample falling between the transformed points of the order statistics. So hereafter, the sample points will be considered to have been transformed and all samples will be considered to be real scalars. Also all observations to be classified will be assumed to have been transformed into scalars. The two classes are assumed to have one region of overlap. For two inseparable classes of samples, the samples of class 2 are taken to lie largely above those of class 1. See Figure II.4 for an example. A decision is made by comparing an unknown observation with two thresholds which are placed in the overlap region.

If the unknown observation z lies above both thresholds, it is assigned to one class; if z lies below both thresholds, it is assigned to the other class; and if z lies between both thresholds, another observation is taken as z lies in the region of overlap. The procedure is applied to the new observation which is compared with a new set of thresholds. It is assumed that all new observations come from the same class. Figure II.5 provides an example of the algorithm showing how the thresholds, labeled A and B, change for each iteration. New observations are taken until a decision is made, and then the algorithm is terminated.

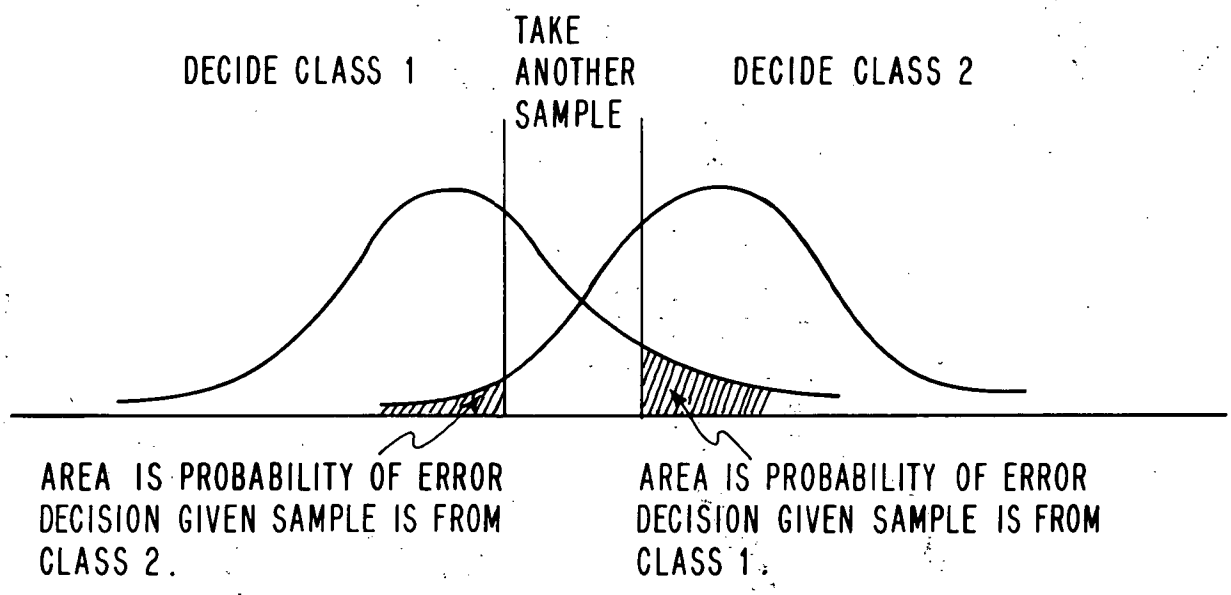


FIG. II.4

Decision Regions for Sequential Test

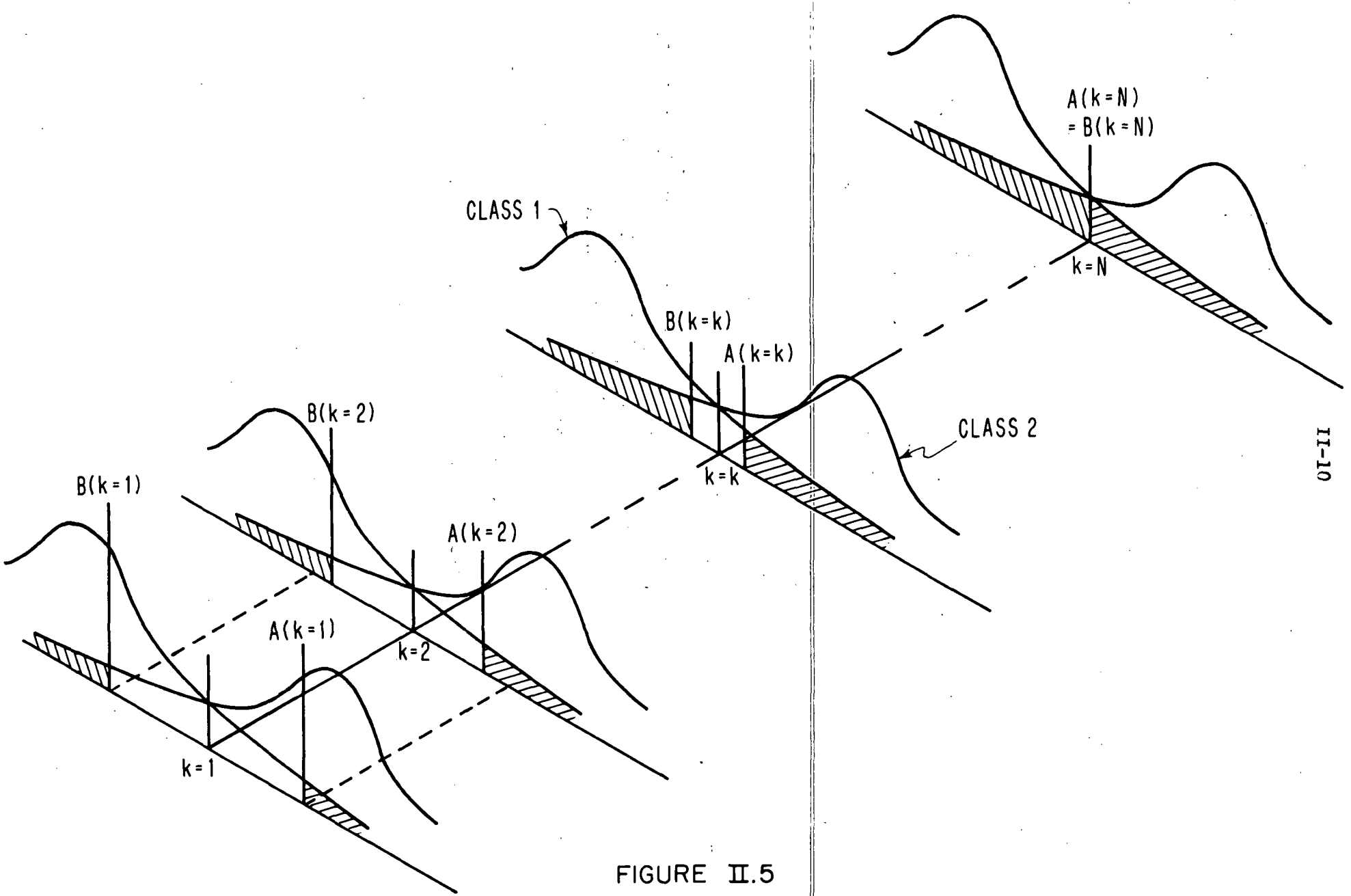


FIGURE II.5

Thresholds Changing in Time

The new observation that is taken in each iteration of the algorithm is compared with a new pair of thresholds that correspond to that iteration.

II.4.2 Setting Thresholds for First Iteration

The thresholds are calculated by using some theory from order statistics on the training sets of each class in such a way as to give an estimate of the probability of a misclassification. The n samples, now scalars, from each of the two training sets are ordered separately in ascending magnitude. The ordering for one class is

$$x_{i_1} < x_{i_2} < \ldots < x_{i_n} \quad .$$

Let the training samples be relabeled for convenience

$$y_1 = x_{i_1},\ y_2 = x_{i_2},\ldots,y_n = x_{i_n} \quad .$$

The training samples are now in ascending order,

$$y_1 < y_2 < \ldots < y_n \quad .$$

If z is an unknown observation, then

$$p(\text{classification error}) = p(\text{classification error} \mid z \in \text{class } 1)p(z \in \text{class } 1) + p(\text{classification error} \mid z \in \text{class } 2)p(z \in \text{class } 2).$$

Thus the error probabilities for each class, p(classification error $|z \in$ class j) j=1,2, can be calculated separately. The setting of thresholds will now be examined in detail for one class, say class 1,

and the ordered training set, $y_1 < y_2 < \ldots < y_n$, will be considered to be from that class. The following discussion of setting thresholds applies to either class.

Given the set of ordered statistics from one class,

$$Y_1 < Y_2 < \ldots < Y_n,$$

the probability that an observation from this class is less than any member of the ordered statistic, Y_j, is $F(Y_j)$. From equation (II.1)

$$E(F(Y_j)) = \frac{j}{n+1} . \qquad \text{(II.6)}$$

An estimated $100j/(n+1)$ percent of all future observations lie below Y_j (or $100(n+1-j)/(n+1)$ percent exceed Y_j.) Figure II.6 gives an example with the two training sets together. The overlap region of the inseparable training sets has been taken to be at the higher end of the class 1 order statistics and lower end of the class 2 order statistics.

In the following formulation of the thresholds, A(k) represents the upper threshold and B(k) the lower threshold where k represents the number of the iteration of the sequential test. A(k=1) will now be determined in such a way that p(classification error$|z \in$ class 1) can be estimated. If the first unknown sample lies above both thresholds, it will be classified as belonging to class 2 which would be an error. If it lies below both thresholds a correct classification of class 1 would be made. If it falls between the thresholds, another observation should be taken. See Figure II.7. To obtain an estimate of the

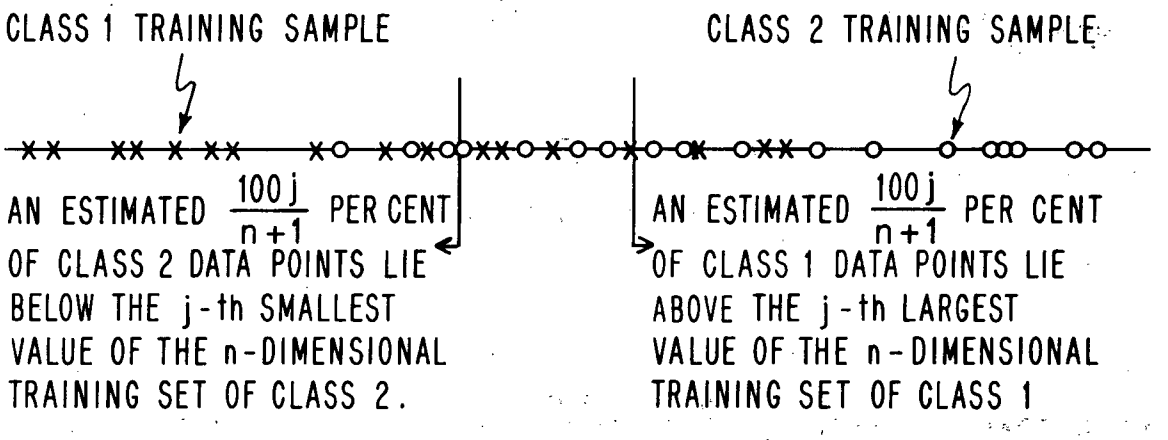


FIG. II.6

Estimating Probabilities from Training Samples

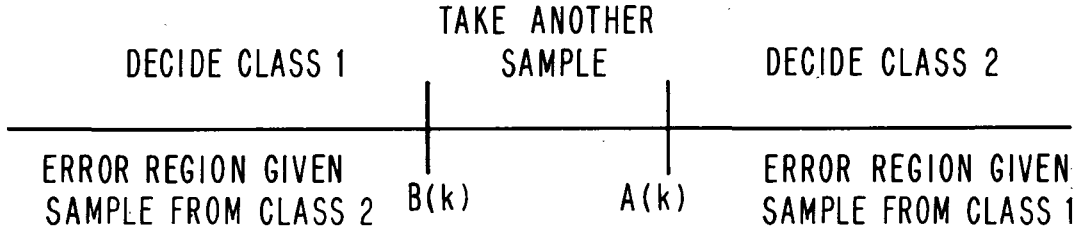


FIG. II.7

Thresholds for Sequential Test

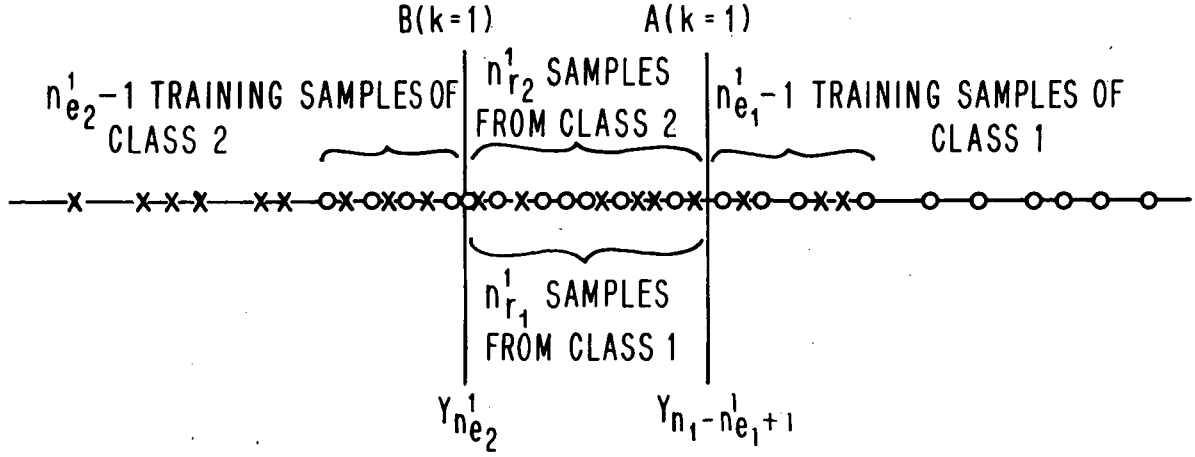


FIG. II.8

Estimating Thresholds for Sequential Test

probability of an observation from class 1 falling above the upper threshold, the number of training samples from class 1 that fall above the threshold A(k=1) can be used.

Let the threshold A(k=1) be set equal to the value of the $n_{e_1}^1$-th largest order statistic of the training set of class 1, $A(k=1) = Y_{n_1-n_{e_1}^1+1}$; then $n_{e_1}^1-1$ training samples lie above A(k=1). The superscript on n represents the number of iterations and the subscript the class. Thus

$$E[p(z_1>A(k=1)\,|\,z_1\varepsilon C^1] = E[1-F(Y_{n_1-n_{e_1}^1+1})] \quad ,$$

and from equation (II.1)

$$E(1-F(Y_{n_1-n_{e_1}^1+1})) = 1-E(F(Y_{n_1-n_{e_1}^1+1}))$$

$$= 1- \frac{n_1-n_{e_1}^1+1}{n_1+1}$$

$$E(1-F(Y_{n_1-n_{e_1}^1+1})) = \frac{n_{e_1}^1}{n_1+1} \qquad \text{(II.7)}$$

If p is the desired probability of error for class 1 on this iteration, then $n_{e_1}^1$ should be chosen so that

$$\frac{n_{e_1}^1}{n_1+1} = p$$

and so solving for $n^1_{e_1}$

$$n^1_{e_1} = (n_1 + 1)p \ . \qquad \text{(II.8)}$$

When $n^1_{e_1}$ is not an integer, the greatest integer less than $n^1_{e_1}$ is used; [w] will represent the largest integer less than or equal to w. A(k=1) is then set equal to $Y_{n_1+1-[n_{e_1}]}$.

B(k=1), the error threshold for class 2, is determined similarly from class 2 training samples. As the error region for class 2 lies at the lower end of the ordered training samples, B(k=1) is set equal to the $n^1_{e_2}$-th lowest order statistic of class 2,

$$E(F(Y_{n^1_{e_2}})) = \frac{n^1_{e_2}}{n_2+1} \ . \qquad \text{(II.9)}$$

$n^1_{e_2}$ is chosen such that p is the desired error probability of an observation from class 2 on the first iteration,

$$\frac{n^1_{e_2}}{n_2+1} = p$$

$$n^1_{e_2} = (n_2 + 1)p \ . \qquad \text{(II.10)}$$

B(k=1) is set equal to $Y_{[n^1_{e_2}]}$. The setting of A(k=1) and B(k=1) is illustrated in Figure II.8.

II.4.3 Thresholds for the Second and Following Iterations

If the observation on the first iteration falls between the thresholds, a second observation is taken. Figure II.5 provides an example. New thresholds are found for testing the second observation. The probability of the first observation falling between the thresholds can be estimated by counting the number of training samples between the thresholds for each class. Again taking class 1, let $n^1_{r_1}$ be the number of training samples between the thresholds on the first iteration, see Figure II.8. Then an estimated $(n^1_{r_1} + 1)/(n_1+1)$ percent of the area under the density function for class 1 falls in the region between the thresholds.

Actually the lower threshold is based on class 2 so that there is not one whole interval between class 1 sample points but a fraction of one at the lower end of the region between the thresholds. In practice, $n^1_{r_1}$ is usually large enough that counting the interval as a whole has a negligible effect on $n^2_{e_1}$.

For a decision to be made resulting in a classification error on the second iteration, the first observation must fall in the region between the thresholds of the first iteration and the second observation in the error region of the second iteration. If p is the desired probability of error for the second iteration, then we desire

p(1st observation between thresholds)p(2nd observation in error region) = p

$$p(B(k=1) < z_1 < A(k=1))p(z_2 > A(k=2)) = p$$

But $p(B(k=1) < z_1 < A(k=1))$ and $p(z_2 > A(k=2))$ are unknown, and they can only be estimated. So the number of training samples in the error region for the second iteration is chosen as

$$\frac{n_{r_1}^1 + 1}{n_1+1} \quad \frac{n_{e_1}^2}{n_1+1} = p$$

$$n_{e_1}^2 = \frac{n_1 + 1}{n_{r_1}^1 + 1} (n_1+1)p$$

$$n_{e_1}^2 = \frac{n_1 + 1}{n_{r_1}^1 + 1} n_{e_1}^1 \qquad \text{(II.11)}$$

from equation (II.9). $A(k=2)$ is set equal to the $[n_{e_1}^2]$-th largest training sample, $A(k=2) = Y_{n_1-[n_{e_1}^2]+1}$. $B(k=2)$ is set similarly using the training samples of class 2. It is desired that $p(B(k=1) < z_1 < A(k=1))p(z_2 < A(k=2)) = p$ which can be estimated by considering

$$\frac{n_{r_2}^1 + 1}{n_2 + 1} \quad \frac{n_{e_2}^2}{n_2+1} = p,$$

and solving for $n_{e_2}^2$

$$n_{e_2}^2 = \frac{n_2+1}{n_{r_2}^2 + 1} (n_2+1)p = \frac{n_2 + 1}{n_{r_2}^1 + 1} n_{e_2}^1 . \qquad \text{(II.12)}$$

$B(k=1)$ is set equal to $Y_{[n_e^2]}$.

As $\frac{n_1+1}{n_{r_1}^1+1} > 1$, then by referring to equation (II.11) it is seen that $n_{e_1}^2 > n_{e_1}^1$ which implies $A(k=2) \leq A(k=1)$, and for similar reasons $B(k=2) \geq B(k=1)$. Thus the thresholds for the second iteration will be closer together than the thresholds for the first iteration.

The number of training samples between the thresholds for the second iteration are counted for each class, $n_{r_1}^2$ and $n_{r_2}^2$. Then $n_{e_1}^3$ and $n_{e_2}^3$ can be calculated. For an error decision on the third iteration both, the first and second observations must fall between their respective thresholds, and the third observation must fall in the error region.

The calculation of the thresholds continues, with the thresholds for each iteration being calculated simultaneously. Figure II.5 again gives an illustrative example. In general,

$$p(B(k=1) < z_1 < A(k=1)) \cdots p(B(k-1) < z_{k-1} < A(k-1)) p(z_k > A(k)) = p$$

The estimated form is

$$\frac{(n_{r_1}^1+1)}{(n_1+1)} \quad \frac{(n_{r_1}^2+1)}{(n_1+1)} \quad \cdots \quad \frac{(n_{r_1}^{k-1}+1)}{(n_1+1)} \quad \frac{n_{e_1}^k}{(n_1+1)} = p, \qquad \text{(II.13)}$$

and solving for $n_{e_1}^k$

$$n_{e_1}^k = \frac{(n_1+1)}{(n_{r_1}^{k-1}+1)} \frac{(n_1+1)}{(n_{r_1}^{k-2}+1)} \cdots \frac{(n_1+1)}{(n_{r_1}^{1}+1)} (n_1+1)p \qquad \text{(II.14)}$$

$$n_{e_1}^k = \frac{(n_1+1)}{(n_{r_1}^{k-1}+1)} n_{e_1}^{k-1} . \qquad \text{(II.15)}$$

Similarly,

$$n_{e_2}^k = \frac{(n_2+1)}{(n_{r_2}^{k-1}+1)} n_{e_2}^{k-1} . \qquad \text{(II.16)}$$

A(k) is set equal $Y_{n_1-[n_{e_1}^k]+1}$ and B(k) equal $Y_{[n_{e_2}^k]}$. As

$\frac{n_1+1}{n_{r_1}^{k-1}+1} > 1$ and $\frac{n_2+1}{n_{r_2}^{k-1}+1} > 1$, the bounds move closer together.

Eventually, for some k, the thresholds will cross. This happens when $n_{e_1}^k$ and $n_{e_2}^k$ become sufficiently large that B(k) > A(k). The algorithm will be terminated for this value of k, and the two thresholds are replaced by a common threshold. Let this terminal value of k be called N. A decision will be made at k = N if the algorithm proceeds this far. In the examples to follow the common threshold was set by averaging the thresholds for k = N-1.

$$A(N) = B(N) = [A(N-1) + B(N-1)]/2 . \qquad \text{(II.17)}$$

This could of course be set in other ways.

The algorithm as presented has taken the probability of error and ending on each iteration to be the same for each class,

$$p(\text{error decision and end on k-th iteration} \mid \text{unknown} \in C^1)$$
$$= p(\text{error decision and end on k-th iteration} \mid \text{unknown} \in C^2) = p.$$

These could be set equal to different values if so desired. Although then the prior probabilities of which class the unknown observation belongs, $p(\text{unknown} \in C^1)$ and $p(\text{unknown} \in C^2)$, should be known in order to calculate the estimated error decision probabilities.

II.5 Application of Algorithm

The application of the algorithm can be divided into two parts, the formulation of the thresholds and the use of the thresholds to classify an unknown observation. This section briefly reviews the steps involved in both parts. Figure II.5 can be referred to as an example.

First the thresholds are set using training sets from the two classes. An ordering function is chosen that separates to some degree the two classes of training samples, and the training samples are reduced to scalars using the ordering function. The training sets of scalars from each class are ordered,

$$\text{Class 1 : } y_1^1 < y_2^1 < \ldots < y_{n_1}^1 \qquad \text{Class 2 : } y_1^2 < y_2^2 < \ldots < y_{n_2}^2$$

The parameter p is chosen. The number of samples in the error region for the first iteration is found,

$$n_{e_1}^1 = (n_1 + 1)p \qquad n_{e_2}^1 = (n_2 + 1)p$$

and the thresholds are set,

$$A(k=1) = Y_{n_1+1-[n_{e_1}^1]} \qquad B(k=1) = Y_{[n_{e_2}^1]} .$$

The number of training samples between B(k=1) and A(k=1) in each class are counted, $n_{r_1}^1$ and $n_{r_2}^1$ respectively. Then for the second iteration, k=2,

$$n_{e_1}^2 = \frac{n_1+1}{n_{r_1}^1+1} \; n_{e_1}^1 \qquad n_{e_2}^2 = \frac{n_2+1}{n_{r_2}^1+1} \; n_{e_2}^1$$

$$A(k=2) = Y_{n_1+1-[n_{e_1}^2]} \qquad B(k=2) = Y_{[n_{e_2}^2]} .$$

Then $n_{r_1}^2$ and $n_{r_2}^2$ are determined by counting the number of training samples of class 1 and class 2 between B(k=2) and A(k=2). For any iteration k,

$$n_{e_1}^k = \frac{n_1+1}{n_{r_1}^{k-1}+1} \; n_{e_1}^{k-1} \qquad n_{e_2}^k = \frac{n_2+1}{n_{r_2}^{k-1}+1} \; n_{e_2}^{k-1}$$

$$A(k=k) = Y_{n_1+1-[n_{e_1}^k]} \qquad B(k=k) = Y_{[n_{e_2}^k]} .$$

Determine $n_{r_1}^k$ and $n_{r_2}^k$ by counting the samples of class 1 and class 2 between B(k) and A(k). Whenever $A(k) \leq B(k)$, call k = N and set one common threshold A(N) = B(N).

In applying the algorithm to classify unknown observations each observation is first reduced to a scalar by using the ordering function. The first observation z_1 is compared to the thresholds A(k=1) and B(k=1). If

$z_1 < B(k=1)$	decide class 1
$z_1 > A(k=1)$	decide class 2
$B(k=1) < z_1 < A(k=1)$	take another observation

If another observation is taken, z_2, then it is similarly compared to A(k=2) and B(k=2). At each iteration that is needed, the bounds for that iteration are used. For any iteration k,

$z_k < B(k)$	decide class 1
$z_k > A(k)$	decide class 2
$B(k) < z_k < A(k)$	take another observation

If the procedure goes until k = N, a decision will be made then as there is only one threshold.

II.6 Estimated Probability of Misclassification

The probability of misclassification for the algorithm will now be considered. The algorithm can end on only one iteration so the events of ending with an error decision on the k-th iteration and of ending with an error decision on the j-th iteration are mutually exclusive for $k \neq j$. The probability of error can be expressed as

$$p(\text{error decision}) = \sum_{k=1}^{N} p(\text{error decision and end on k-th iteration}) \tag{II.18}$$

where

$$\begin{aligned} &p(\text{error decision and end on k-th iteration}) \\ &\quad = p(\text{error decision and end on k-th iteration} \mid \text{unknown} \in C^1) \\ &\qquad \cdot p(\text{unknown} \in C^1) \\ &\quad + p(\text{error decision and end on k-th iteration} \mid \text{unknown} \in C^2) \\ &\qquad \cdot p(\text{unknown} \in C^2). \end{aligned} \tag{II.19}$$

Consider first the case where the unknown observations $z_1, z_2, \ldots, z_k$ are from class 1. Then

$$\begin{aligned} &p(\text{error decision and end on k-th iteration} \mid \text{unknown} \in C^1) \\ &\quad = p(B(1) < z_1 < A(1))p(B(2) < z_2 < A(2)) \cdots \\ &\qquad p(B(k-1) < z_{k-1} < A(k-1))p(z_k > A(k)). \end{aligned} \tag{II.20}$$

All the thresholds are calculated from the training samples, and so

$$p(B(1) < z_1 < A(1)), p(B(2) < z_2 < A(2)), \ldots, p(B(k-1) < z_{k-1} < A(k-1), \; p(z_k > A(k))$$

are random variables. Also since the thresholds were calculated from the same training samples, these random variables are dependent, and the expectation of the left hand side of equation (II.20) is not equal to the product of the expectations of the terms on the right hand side. As it is not readily apparent how the true expectation

can be found, the expectation is approximated, however, by

$$\hat{p}(\text{error decision and end on k-th iteration} \mid \text{unknown} \in C^1)$$

$$= Ep(B(1) < z_1 < A(1))Ep(B(2) < z_2 < A(2))\cdots$$

$$Ep(B(k-1) < z_{k-1} < A(k-1))Ep(z_k > A(k)). \qquad \text{(II.21)}$$

The symbol $\hat{p}$ is used to denote that the term is an approximation of the expected value.

By the construction of the algorithm,

$$\hat{p}(\text{error decision and end on k-th iteration} \mid \text{unknown} \in C^1) = p. \qquad \text{(II.22)}$$

A similar procedure can be used to show

$$\hat{p}(\text{error decision and end on k-th iteration} \mid \text{unknown} \in C^2) = p. \qquad \text{(II.23)}$$

Thus from equation (II.19),

$$\hat{p}(\text{error decision and end on k-th iteration})$$

$$= p \cdot p(\text{unknown} \in C^1) + p \cdot p(\text{unknown} \in C^2) = p,$$

and so

$$\hat{p}(\text{error decision}) = \sum_{k=1}^{N} p$$

$$\hat{p}(\text{error decision}) = Np.^{*} \tag{II.24}$$

As mentioned previously, Np is not the true expected probability of error since the product of expectations of dependent random variables was taken. Loosely speaking if there are one hundred training samples, the addition of another sample provides more information to revise the estimate of the probability of error than if there are one thousand samples. Thus as the number of

---

*Since p is a specified parameter, it can be shown that $Np \leq 1$ by showing that N, the maximum number of iterations, has an upper bound of 1/p. N will have its largest value when the probabilities of an observation falling between the thresholds and not being classified at each iteration have their largest values. Consider first the probability of an error decision given the string of observations is from class 1. At each iteration, $Ep(z_k \geq A(k))$ is determined before $Ep(B(k) < z_k < A(k))$ is determined, and thus the upper bound on $Ep(B(k) < z_k < A(k))$ is $1-Ep(z_k \geq A(k))$. For convenience, let $p_{ek} = Ep(z_k \geq A(k))$ and so $1-p_{ek}$ is the upper bound on $Ep(B(k) < z_k < A(k))$. Using these upper bounds, the thresholds at each iteration are found by setting

$$(1-p_{e1})(1-p_{e2})\cdots(1-p_{e(k-1)})p_{ek} = p$$

as is done in equations (II.13) and (II.14). For $k = 1$, the thresholds are set such that $p_{e1} = p$, and by induction, it can be shown that $p_{ek} = p/[1-(k-1)p]$ when the above equation is used to determine the thresholds. The thresholds are determined so that the fraction of training samples exceeding A(k) is equal to p_{ek}. Since the fraction cannot exceed one, the procedure for generating the thresholds at each iteration will stop before p_{ek} equals 1. Thus $p_{ek} = p[1-(k-1)p] \leq 1$ which implies $k \leq 1/p$. The analysis is similar when the string of observations is assumed to be from class 2, and the same upper bound on k is found. Thus $N \leq 1/p$ and $\hat{p}(\text{error decision}) \leq 1$.

training samples approaches infinity, the knowledge of $p(B(k) < z_k < A(k))$, $k=1,2,\ldots,N$, becomes precise and the bias in the estimates of the probability of error would be expected to tend to zero. Also in the next section, a comparison is made of experimental results of the algorithm trained on one set of training samples with results of using a different set of training samples to calculate the pair of thresholds at each iteration. The use of a different set of training samples to calculate the pair of thresholds at each iteration makes the terms $p(B(k) < z_k < A(k))$, $k=1,2,\ldots,N$, independent so Np is actually the expected probability of error. In most practical problems, however, using a different set of training samples at each iteration would require an excessive number of training samples. The experimental comparison showed there was little effect on the experimental results of using the same set of training samples. A slight approximation was also introduced when the value calculated for the number of a training sample was not an integer and the largest integer less than the value was used. These approximations seem unavoidable when the number of training samples is finite.

If Np is not near the desired value, p can be varied, which will change N and hence Np. N is dependent on the value of p chosen, and generally for smaller p, N becomes larger. N is also dependent on the two sets of training samples. If the training sets have a large overlap, N will be large. This is to be expected as the region of indecision is large so more iterations will result.

The probability of making an error decision on the N-th or last iteration is actually not equal to p as the two thresholds were combined into one instead of allowing them to cross. The actual error estimate can be made by counting the number of training samples for class 1 and class 2 which would result in an error decision on the N-th iterations. Let $m_{e_1}^N$ be the number of training samples of class 1 above A(N) = B(N) and $m_{e_2}^N$ be the number of training samples of class 2 below. Then

$$\hat{p}(\text{error decision and end on N-th iteration} \mid \text{unknown } \epsilon \text{ class 1})$$

$$= \frac{n_{r_1}^1 + 1}{n_1 + 1} \cdot \frac{n_{r_1}^2 + 1}{n_1 + 1} \cdots \frac{n_{r_1}^{N-1} + 1}{n_1 + 1} \cdot \frac{m_{e_1}^N}{n_1 + 1}$$

$$= m_{e_1}^N \frac{p^N}{n_{e_1}^N}$$

from equation (II.14) where $n_{e_1}^N$ is defined by equation (II.14). A similar equation applies to class 2. The total estimated probability of error is

$$\hat{p}(\text{error decision}) = (N-1)p + m_{e_1}^N \frac{p^N}{n_{e_1}^N} p(\text{unknown } \epsilon \text{ class 1}) + m_{e_2}^N \frac{p^N}{n_{e_2}^N} p(\text{unknown } \epsilon \text{ class 2})$$

As p is small, Np gives an adequate expression for $\hat{p}$(error decision) for most values of N and p.

An intuitive explanation for the closing together of the thresholds

can be given. In order for the algorithm to proceed to the second iteration, the first observation must fall between the first two thresholds. For a decision to be made resulting in an error on the second iteration, the second observation must fall in the error region. Let p be the desired probability of making a decision which ends in an error at each iteration. To obtain p on the first iteration, the probability of falling in the error region should be p. For an error decision to be made on the second iteration, the first observation must fall between the thresholds and the second observation in the error region. The probability of this is $p(B(k=1) < z_1 < A(k=1))$ $\cdot$ $p(z_2 \in$ error region for $k=2) = p$. As $p(B(k=1) < z_1 < A(k=1)) < 1$, $p(z_2 \in$ error region for $k=2)$ is greater than $p(z_1 \in$ error region for $k=1)$, and thus the error decision region for k=2 can afford to be larger than for k=1 leading to a smaller overlap region. The same argument applies for larger k.

The setting of the estimated p(error decision on iteration k) equal to p for each iteration was done so that an estimate of the probability of error for the algorithm could be obtained. This also resulted in a finite number of iterations for the algorithm. The probability of error is estimated looking from the beginning of the test before any samples are taken.

II.7 Remarks

In treating multidimensional samples in the experimental results of the next section, the same linear ordering function was used in

determining the thresholds for all the iterations. Using the same linear ordering function throughout the algorithm may be suitable when the data comes from unimodal densities which have one region of overlap between the two classes. For some sample densities, another type of ordering function might be preferable. The most desirable procedure would be not only to locate a plane for each threshold, but to determine the orientation of the plane in order to optimize the procedure. At each iteration, all coefficients $\{\alpha_i\}$ in the linear ordering function $\alpha_1 x^1+\alpha_2 x^2+\ldots+\alpha_s x^s = \alpha_o$ would be determined instead of finding only α_o. For example, the number of training samples to be placed in the error region for each threshold, $n_{e_1}^k$ and $n_{e_2}^k$, could be found as explained previously. For each iteration, a plane would be placed through a training sample of class 1 so that $n_{e_1}^k$ samples from class 1 lay on the error decision side of the plane and the plane oriented so that the number of training samples of class 2 on the same side was maximized. Such a technique would set $\hat{p}$(error decision on k-th iteration|class 1) = p and maximize the probability of a correct classification for a class 2 observation. A similar procedure would be applied using class 2 training samples to the other plane and the class 2 error region. This method would give $\hat{p}$(error decision on k-th iteration|class i)= p, i=1,2, and would also minimize the number of iterations. But this technique requires a considerable amount of computation. Such a procedure might have to be repeated several times to find the best training

sample through which to place the plane, and then the computation must be done for the planes at each iteration. The choice of an ordering function for multidimensional sample pattern classification is an area in which further work can be done. Of course for scalar samples the question of choice of an ordering function does not occur. For the examples in the next section, a single linear ordering function was thought to be sufficient considering the extra amount of computation required to orient a different plane at each iteration.

II.8 Experimental Results

The algorithm was tested on Gaussian random variables and on electroencephalogram (EEG) signals. The results for scalar Gaussian samples are given in Table II.1. Several training set sizes and several values of the parameter p are given. The algorithm for each set of parameters was tested on one thousand observations from each of the two classes.

The algorithm was also tested on EEG signals which are discussed in Section I.2 and in Appendix II.2. The EEG signals are from a subject with a strobe light flashing in his eye or from the subject with the light off. It is desired to decide on the basis of EEG signals if the light is flashing or not. The signals with the light off will be called class 1 and with the light on class 2. The EEG responses were continuous signals of 100 millisecond duration, and the responses were sampled every millisecond to obtain a one hundred dimensional vector for each sample. A feature reduction scheme of

Parameters		Experimental Results					
p	Number of training samples for each class $n_1 = n_2$	N = maximum number of iterations for decision	Average number of experimental iterations for decision		Estimated error rate = Np	Class 1 mean=-.8 experimental error rate	Class 2 mean= .8 experimental error rate
			Class 1	Class 2			
p = .01	99	12	4.74	4.54	.12	.0474	.0666
p = .01	199	9	4.03	3.95	.09	.0444	.0712
p = .01	399	9	3.93	3.70	.09	.0630	.0741
p = .01	999	7	3.3	2.90	.07	.11	.058
p = .005	199	13	4.95	4.74	.065	.0346	.0711
p = .005	399	13	5.02	5.12	.065	.0352	.0718
p = .005	999	10	4.33	3.89	.05	.065	.055

Variance of both classes = 1

TABLE II.1

Gaussian Experimental Error Rates

Prabhu [1, 8], which is explained in Appendix II.1, was used to select a smaller number of features from the 100-dimensional vector to make the testing procedure more manageable. For most tests two features were used. Of the one hundred features, features eighty-five and fifty-seven were selected as containing the most significant information. A linear ordering function was used,

$$y = \alpha_{57}x_{57} + \alpha_{85}x_{85}$$

The algorithm was trained on one section of EEG data from the subject and tested on another section from the same subject. Table II.2 gives error rates on the testing samples for several parameter p values. The samples were taken serially as they appeared from the patient. Five hundred testing observations were used in all cases.

An examination of the EEG responses showed that the samples are correlated and nonstationary. The independence assumption of the algorithm is violated. The nonstationarity means that the samples are not identically distributed. The correlation of the samples along with the nonstationarity contributed to the higher than estimated error rates in Table II.2.

To test the algorithm on data which was independent and uncorrelated, one thousand serial samples of EEG waveforms were mixed together so they no longer appeared serially as they were recorded from the patient. The results for the mixed samples appear in Table II.3. The experimental error rates in this case agree more closely with the estimated error rates. This indicated that all the assumptions of the algorithm are

Parameters		Experimental Results					
p	Number of training samples for each class $n_1 = n_2$	N = maximum number of iterations for decision	Average number of experimental iterations for decision		Estimated error rate = Np	Class 1 (no strobe) experimental error rate	Class 2 (strobe on) experimental error rate
			Class 1	Class 2			
p = .01	99	6	1.9	1.8	.06	.209	.0757
p = .01	199	7	2.22	2.55	.07	.186	.0612
p = .01	399	8	3.42	3.05	.08	.199	.0548
p = .01*	999	9	3.57	4.13	.09	.128	.066
p = .005	199	12	3.68	6.25	.06	.11	.0875
p = .005	399	11	3.91	4.38	.055	.132	.0789
p = .005*	999	14	4.8	6.10	.07	.107	.013
p = .001*	999	40	13.9	20.8	.04	.0556	.0833

* Five features instead of two were used for these experiments.

TABLE II.2

EEG Experimental Error Rates

Parameters		Experimental Results					
p	Number of training samples for each class $n_1 = n_2$	N = maximum number of iterations for decision	Average number of experimental iterations for decision		Estimated error rate = Np	Class 1 (no strobe) experimental error rate	Class 2 (strobe on) experimental error rate
			Class 1	Class 2			
p = .01	99	10	4.81	5.15	.1	.0962	.103
p = .01	199	10	4.63	5.88	.1	.0925	.1295
p = .01	399	10	3.68	4.58	.1	.054	.11
p = .005	199	16	5.95	8.33	.08	.0357	.12
p = .005	399	15	5.05	7.58	.075	.0303	.166

TABLE II.3

Independent EEG Experimental Error Rates

not met by the EEG waveforms as they are recorded from the a patient.

Section II.6 mentioned that the estimated probability of error Np is biased since all the thresholds are calculated from the same training samples. Table II.4 shows a comparison of experimental results of the algorithm trained on one set of training samples with the results of using a different set of training samples to calculate the pair of thresholds at each iteration. The examples are Gaussian as appear in Table II.1, and p = .01 was used for all the results.

	Number training samples in each class	Estimated error = Np	Class 1 experimental error rate	Class 2 experimental error rate
One training set	99	.12	.0474	.0666
Different training sets		.09	.0468	.0675
One training set	199	.09	.0444	.0712
Different training sets		.08	.0947	.0655

TABLE II.4 Comparison of Error Rates for One Training Set vs Several Training Sets

The table indicates that using a different set of training samples for calculating the pair of thresholds at each iteration does not give significantly different experimental results than using one set of training samples. The difference between the two estimated error rates

decreased as the number of training samples increased.

II.9 Conclusion to Chapter II

The algorithm presented in this chapter is a sequential approach to pattern classification for the case where the underlying probability densities of each class are unknown but training sets are available. When a linear ordering function is used, the algorithm can be viewed as a sequential variation of the linear separating plane approach to pattern classification. The algorithm used a different pair of thresholds at each iteration of the sequential test. The thresholds are calculated before the test and are independent of the observations taken during the sequential decision procedure. The method does require some prior assumptions on the pattern classes. The classes should have one region of overlap such that when the multidimensional samples of the two classes are transformed to scalars the new scalar samples of one class lie largely below the new scalar samples of the other class. For example if one class of samples is surrounded by samples of the other, the classes cannot be separated by a linear transformation. A nonlinear transformation would have to be found.

The algorithm presented in this chapter used a different pair of thresholds at each iteration of the sequential test, the next few chapters present a sequential test where the same pair of thresholds is used throughout the test.

Appendix II.1 - Feature Reduction and Separating Hyperplanes

The feature reduction scheme used in this report for selecting significant features out of a vector random sample of many features was developed by Prabhu [1], [8]. A measure of effectiveness of any particular feature for classification purposes is

$$\frac{[\mu_i^1 - \mu_i^2]^2}{\sigma_{ii}^1 + \sigma_{ii}^2} \qquad \text{(II.1.1)}$$

where μ_i^j and σ_{ii}^j are the mean and variance of the i-th feature of class j. The criterion picks the feature that tends to maximize the distance between the means of the two classes while minimizing the dispersion about the means. Considering the combined effectiveness of a group of features, the correlation between the features is taken into account, and the criterion generalizes to

$$d = (\mu^1 - \mu^2)^T(\Sigma^1 + \Sigma^2)^{-1}(\mu^1 - \mu^2) \qquad \text{(II.1.2)}$$

where μ^j and Σ^j are the mean vector and covariance matrix of the features under consideration from class j. Since the means and covariances of the two classes are unknown for the examples considered in this thesis, the means and covariances are estimated from training sets of the two classes.

Let d_m be the value of the criterion in equation (II.1.2) when m features are considered. The algorithm for selecting features from a vector of s features, $x = (x^1, x^2, \ldots, x^s)$ is :

1.) Select the first feature x^i such that

$$\frac{(\mu_i^1 - \mu_i^2)^2}{\sigma_{ii}^1 + \sigma_{ii}^2} = \max_j \frac{(\mu_j^1 - \mu_j^2)^2}{\sigma_{jj}^1 + \sigma_{jj}^2}$$

and so

$$d_1 = \frac{(\mu_i^1 - \mu_i^2)^2}{\sigma_{ii}^1 + \sigma_{ii}^2} \quad .$$

2.) At each subsequent step after m features have been chosen and d_m calculated, the increase in the criterion $(d_{m+1}-d_m)$ is computed for each of the remaining features. The feature that gives rise to the maximum increase is chosen.

Thus the algorithm at each step selects the feature that adds the most to the effectiveness of the feature set already chosen where the effectiveness is measured by equation (II.1.2). The feature selection procedure is not truely optimal in that the subset of the best m features is not necessarily a subset of the best m+1 features. To be truely optimal, the algorithm must search over all possible combinations of m features at each step. But such an exhaustive search becomes quickly infeasible as the total number of features increases.

The separating hyperplane that was used for transforming vector samples into scalars in this report is

$$\alpha_0^T x + \beta_0 = 0$$

where

$$\alpha_o = (\Sigma^1 + \Sigma^2)^T(\mu^1 - \mu^2)$$

$$\beta_o = -\frac{1}{2}(\mu^1 - \mu^2)^T(\Sigma^1 + \Sigma^2)^{-1}(\mu^1 - \mu^2) \quad . \qquad \text{(II.1.3)}$$

The weighting vector α_o maximizes

$$\frac{[\alpha^T(\mu^1 - \mu^2)]^2}{\alpha^T(\Sigma^1 + \Sigma^2)\alpha}$$

which is interpreted as the ratio of the distance between the means of the classes to the dispersion of the classes along the direction α. If the classes are Gaussian, $N(\mu^1,\Sigma^1)$ and $N(\mu^2,\Sigma^2)$ respectively, then $\alpha_o^T x + \beta_o$ is the separating surface that minimizes the probability of misclassification with the prior probabilities of each class being equal.

Appendix II.2 - EEG Data

A detailed discussion of the EEG data is given by Prabhu [1], and much of the description presented in this appendix is based on Prabhu's discussion. An electroencephalogram (EEG) is a recording of electrical activity of the brain. The electrical activity is of the order of microvolts and is measured by electrodes placed on the surface of the scalp. While the precise origins of the electrical potentials is not yet fully understood, it is generally agreed that the potentials result from the synchronous activity of a large number of cells. In order to maintain some uniformity in the EEG measurements, it is necessary to keep the patient in the same psychological state during different recordings. When the recording is made from an alert patient in a darkened, soundless room cut off from external stimuli, the EEG is said to be "spontaneous."

Since an EEG recording is the result of the combined activity of many cells, an EEG signal can be considered to be a sample from a random process. An example of an EEG is shown in Figure II.9.* The EEG has been observed to have several dominant frequencies with the most dominant between 8.5 and 10.5 c.p.s. This is called the alpha frequency. An EEG record can be split into equal parts where the length of each part is equal to the period of the alpha frequency. The dotted line in Figure II.10 shows the average signal that results

---

*Figures II.9, II.10, and II.11 have been taken from Prabhu [1].

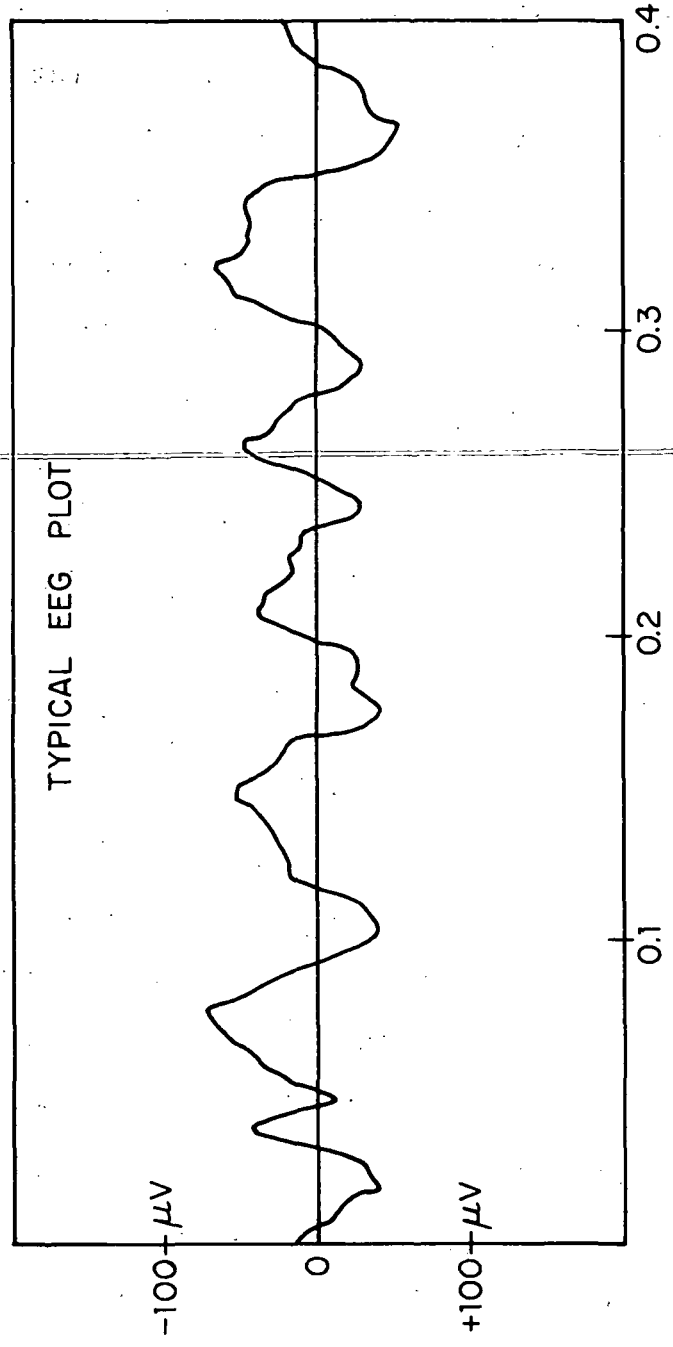


FIGURE II.9

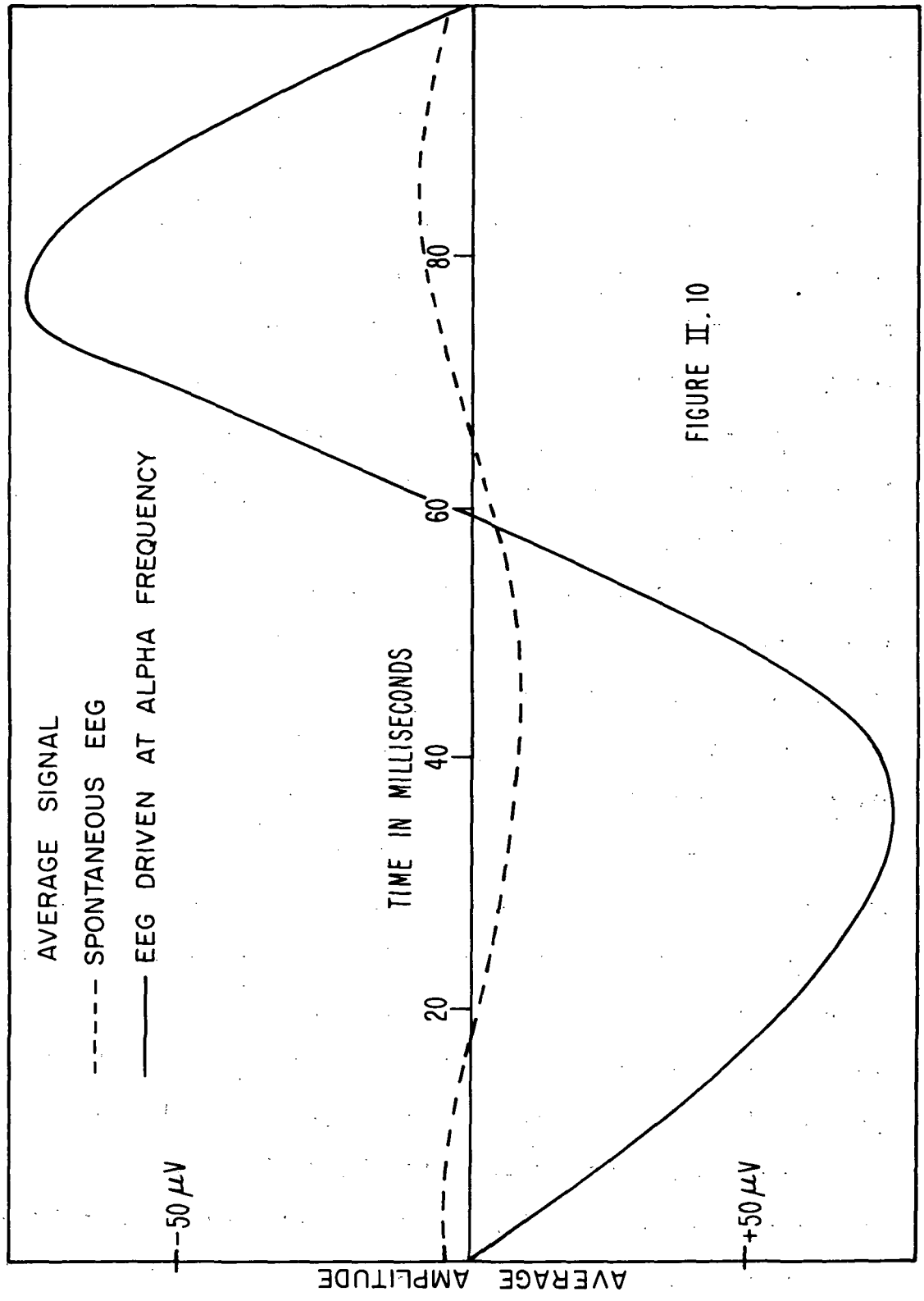

AVERAGE SIGNAL
SPONTANEOUS EEG
EEG DRIVEN AT ALPHA FREQUENCY
TIME IN MILLISECONDS
20
40
60
80
-50 μV
+50 μV
AVERAGE AMPLITUDE

FIGURE II.10

from averaging these parts.

While the spontaneous EEG represents the electrical activity of the brain when no visual or auditory stimuli is present, a different EEG signal can be produced by a flash of light into the patient's eyes through closed eyelids. If a light is flashed periodically at a frequency very near the alpha frequency, then the EEG has the affect of being driven into resonance. The EEG signal between two consecutive flashes is called an "evoked" response, and the solid line in Figure II.10 shows the average signal of the evoked responses.

The classification algorithms tested in this thesis attempted to distinguish between spontaneous EEG and evoked EEG. A signal over one period of the alpha frequency was taken to be one sample.

The EEG record used in this thesis was obtained from NASA through the former Electronics Research Center, Cambridge, Massachusetts. A recording of ten minutes duration was done on a single person in one sitting from a pair of electrodes located in the left occipital-parietal area. Both spontaneous and evoked responses were obtained in the one recording. A stroboscopic light was flashed into the eye of the subject through closed eyelids. The frequency of the flashing was tuned to his alpha rhythm which was approximately 10 c.p.s., and thus a flash occurred every 100 milliseconds. The stroboscopic light was blocked periodically from the eye of the subject, and thereby giving rise to spontaneous EEG. Thus the entire EEG record was composed of blocks of evoked EEG driven at the alpha frequency and of

spontaneous EEG. The length of each block was about 25 seconds.

To facilitate digital computer work, each of the waveforms was discretized by sampling the amplitude every millisecond. Thus each response between two successive stroboscopic stimuli would be expected to have 100 sampled values. In practice, it was found that the number sometimes exceeded 100 due to drifts in the stroboscopic frequency. In order for the pattern vectors to be of uniform dimension, only the first 100 values were retained.

In the experimental work of this thesis, only a few of the 100 features in each digitized waveform were used. The features were selected by the feature reduction procedure explained in Appendix II.1. In order to illustrate the degree of overlap between the two classes of EEG signals, Figure II.11 shows a plot of samples from the two types of EEG. The samples are two dimensional with the features being the first two selected by the feature reduction procedure. The line in the figure is the separating plane for the two features where the equation of the plane is also explained in Appendix II.1. Prabhu [1] found that there was about 20% error rate in classification decisions made on single observations with the separating plane.

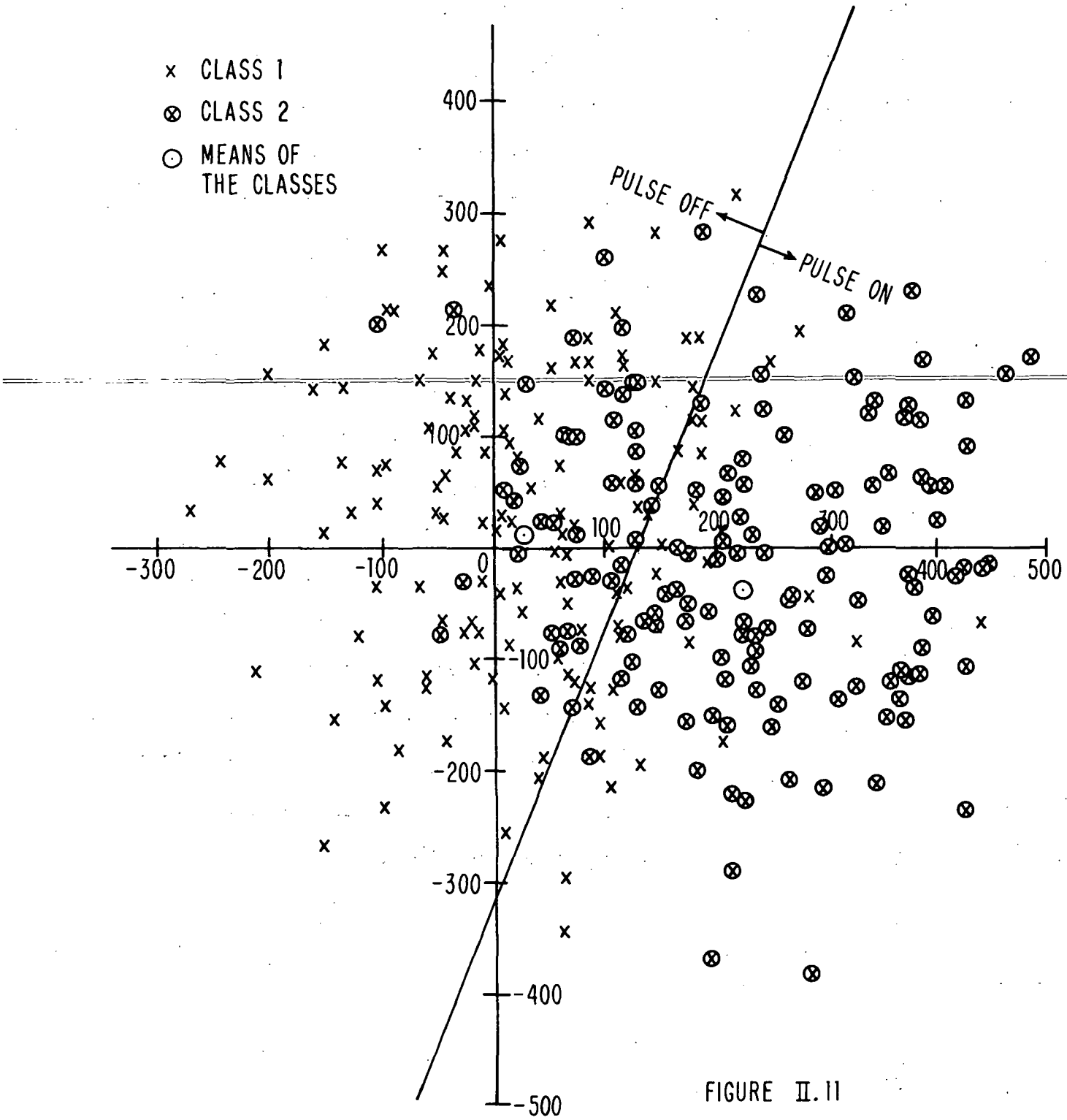


FIGURE II.11

Overlap Between Two Classes of EEG Parrerns

Appendix II.3 - Order Statistics

This appendix will define the notion of an order statistic and present some of the properties of such a statistic. Some references that can be consulted on order statistics are Hogg and Craig [9], Wilks [10], Fraser [11], and David [12].

Let $X_1, X_2, \ldots, X_n$ be n independent random variables identically distributed with absolutely continuous distribution function F(x) and with probability density function f(x). Rearrange $X_1, X_2, \ldots, X_n$ in ascending order so that $X_{i_1} \leq X_{i_2} \leq \cdots \leq X_{i_n}$. For convenience relabel the set as $Y_1 = X_{i_1}$, $Y_2 = X_{i_2}, \ldots, Y_n = X_{i_n}$ so that $Y_1 \leq Y_2 \leq \cdots \leq Y_n$. Y_i, i=1,2,...,n, is called the i-th order statistic of the random sample $X_1, X_2, \ldots, X_n$.

The joint density function of $Y_1, Y_2, \ldots, Y_n$ can be shown to be

$$g(y_1, y_2, \ldots, y_n) = \begin{cases} n! f(y_1) f(y_2) \cdots f(y_n) \\ \qquad y_1 \leq y_2 \leq \cdots \leq y_n \\ 0 \qquad \text{elsewhere} \end{cases} \qquad \text{(II.3.1)}$$

From this joint density, it follows that the marginal probability density function of y_k is

$$g_k(y_k) = \frac{n!}{(k-1)!(n-k)!} [F(y_k)]^{k-1} [1-F(y_k)]^{n-k} f(y_k) , \qquad \text{(II.3.2)}$$

and the joint density of y_i and y_j, $i < j$, is

$$g_{ij}(y_i,y_j)=\begin{cases}\dfrac{n!}{(i-1)!(j-i-1)!(n-j)!}\,[F(y_i)]^{i-1}[F(y_j)-F(y_i)]^{j-i-1} \\ \qquad\cdot[1-F(y_j)]^{n-j}f(y_i)f(y_j) & y_i \leq y_j \\ 0 & \text{elsewhere .}\end{cases} \tag{II.3.3}$$

The distribution function of F(x) will now be considered. Let X be a random variable having an absolutely continuous distribution function F(x) and probability density function f(x). Then the random variable Z = F(X) has a uniform distribution on the interval (0,1). This will be shown under the assumption that f(x) is positive and continuous for a < x < b and zero elsewhere. The distribution function of X can be written as

$$F(x) = \begin{cases} 0 & x \leq a \\ \int_a^x f(u)du & a < x < b \\ 1 & x \geq b \ . \end{cases}$$

Then for the transformation z = F(x), dz/dx = f(x) for a < x < b, and

$$f(x)\left|\frac{dx}{dz}\right| = f(x)\,\frac{dx}{dz} = f(x)\,\frac{1}{dz/dx} = f(x)\,\frac{1}{f(x)} = 1 \quad \text{for } a < x < b \ .$$

Thus the probability density function of Z = F(X) is

$$h(z) = \begin{cases} 1 & 0 < z < 1 \\ 0 & \text{elsewhere} \end{cases} . \qquad \text{(II.3.4)}$$

Since $Z = F(X)$ is a random variable with a uniform distribution on the interval (0,1), it follows that $p(F(X) \leq \nu) = \nu$. That is, if p is the probability that a future sample will fall below the random variable X, then the probability that p does not exceed ν is ν.

Consider again the random sample $X_1, X_2, \ldots, X_n$ and the set of order statistics for this random sample $Y_1, Y_2, \ldots, Y_n$. Consider further the set of random variables $F(X_1), F(X_2), \ldots, F(X_n)$. Since F(x) is nondecreasing in x, it follows that $F(Y_1) \leq F(Y_2) \leq \ldots \leq F(Y_n)$, and hence $Z_1 = F(Y_1), Z_2 = F(Y_2), \ldots, Z_n = F(Y_n)$ are the order statistics of the random sample $F(X_1), F(X_2), \ldots, F(X_n)$. Since F(X) is uniform on the interval (0,1), the joint density function of $Z_1, Z_2, \ldots, Z_n$ is found from equation (II.3.1) to be

$$h(z_1, z_2, \ldots, z_n) = \begin{cases} n! & 0 \leq z_1 \leq z_2 \leq \cdots \leq z_n \leq 1 \\ 0 & \text{elsewhere} \end{cases} \qquad \text{(II.3.5)}$$

Similarly, the marginal density of $Z_k = F(Y_k)$ and the joint density of $Z_i = F(Y_i)$ and $Z_j = F(Y_j)$, $i < j$, can be found from equations (II.3.2) and (II.3.3)

$$h_k(z_k) = \begin{cases} \dfrac{n!}{(k-1)!(n-k)!} z_k^{k-1} (1-z_k)^{n-k} & 0 < z_k < 1 \\ 0 & \text{elsewhere} \end{cases} \qquad \text{(II.3.6)}$$

$$h_{ij}(z_i,z_j) = \begin{cases} \dfrac{n!}{(i-1)!(j-i-1)!(n-j)!}\, z_i^{i-1}(z_j-z_i)^{j-i-1}(1-z_j)^{n-j} & \\ & 0 < z_i < z_j < 1 \\ 0 & \text{elsewhere} \end{cases} . \tag{II.3.7}$$

For the order statistics $y_1,y_2,\ldots,y_n$, the intervals $(-\infty,y_1]$, $(y_1,y_2],\ldots,(y_n,+\infty)$ are called sample blocks. The probabilities of a future observation falling in each of these sample blocks are $F(y_1),F(y_2)-F(y_1),\ldots,1-F(y_n)$ respectively. $F(y_j)-F(y_{j-1})$ is called a coverage of the sample block $(y_{j-1},y_j]$. The distribution of the random variable $Z_j-Z_i = F(Y_j)-F(Y_i)$, $i < j$, will now be considered. It can be shown that the random variable Z_j-Z_i has the same distribution as the random variable Z_{j-i}. Thus from equation (II.3.2), $Z_j-Z_i = F(Y_j)-F(Y_i)$ has the probability density function

$$h(v) = \begin{cases} \dfrac{n!}{(j-i-1)!(n-j+i)!}\, v^{j-i-1}(1-v)^{n-j+i} & 0 < v < 1 \\ 0 & \text{elsewhere} \end{cases} \tag{II.3.8}$$

It is noted that this is a Beta distribution $B(j-i,n-j+i+1)$. The mean and variance of $F(Y_j)-F(Y_i)$, $i < j$, can be calculated to be

$$E[F(Y_j)-F(Y_i)] = \frac{j-i}{n+1} \tag{II.3.9}$$

$$\mathrm{Var}[F(Y_j)-F(Y_i)] = \frac{(j-i)(n-j+i)}{(n+1)^2(n+2)} . \tag{II.3.10}$$

In particular, $E[F(Y_{j+1})-F(Y_j)] = 1/(n+1)$. Thus the order statistics partition the sample axis into n+1 parts, and the expected probability of a future observation falling in each part is $1/(n+1)$.

The theory of sample blocks and coverages can be extended to more than one dimension by using ordering functions. The concept of ordering functions will be introduced by considering a single ordering function to partition the s-dimensional sample space. Let $(X_j^1,X_j^2,\ldots,X_j^s)$, $j=1,2,\ldots,n$, be n independent s-dimensional random variables distributed as the random variable $X = (X^1,X^2,\ldots,X^s)$ with a continuous s-variate distribution function $F(x^1,x^2,\ldots,x^s)$. If $W = t(X^1,X^2,\ldots,X^s)$ is a random variable with a continuous distribution $T(w)$, then $t(x^1,x^2,\ldots,x^s)$ is an ordering funetion. $W_j = t(X_j^1,X_j^2,\ldots,X_i^s)$, $j=1,2,\ldots,n$, constitutes a random sample from a population whose distribution function is $T(w)$, and the random sample can be ordered. Let the order statistics for the random sample $(W_1,W_2,\ldots,W_n)$ be $(W_{i_1},W_{i_2},\ldots,W_{i_n})$. Then the j-th sample block is $B_j = \{x|t(x_{i_{j-1}}) < t(x) \leq t(x_{i_j})\}$ where x_{i_j} is the s-dimensional sample such that $w_{i_j} = t(x_{i_j})$. Figure II.12 provides an illustration in two dimensions. The coverages of the n+1 sample blocks are $Z_1 = T(W_{i_1}), Z_2 = T(W_{i_2})-T(W_{i_1}),\ldots,$ $Z_n = T(W_{i_n})-T(W_{i_{n-1}}), Z_{n+1} = 1-T(W_{i_n})$ where $Z_j = T(W_{i_j})-T(W_{i_{j-1}})$ is the probability that a future observation will fall in the j-th sample block. It can be shown that for the coverages $Z_1,Z_2,\ldots,Z_{n+1}$ the sum of any r coverages has a Beta distribution $B(r,n+1-r)$. Thus the expected value of a future observation falling in any r, $r \leq n$, of the sample blocks is $r/(n+1)$.

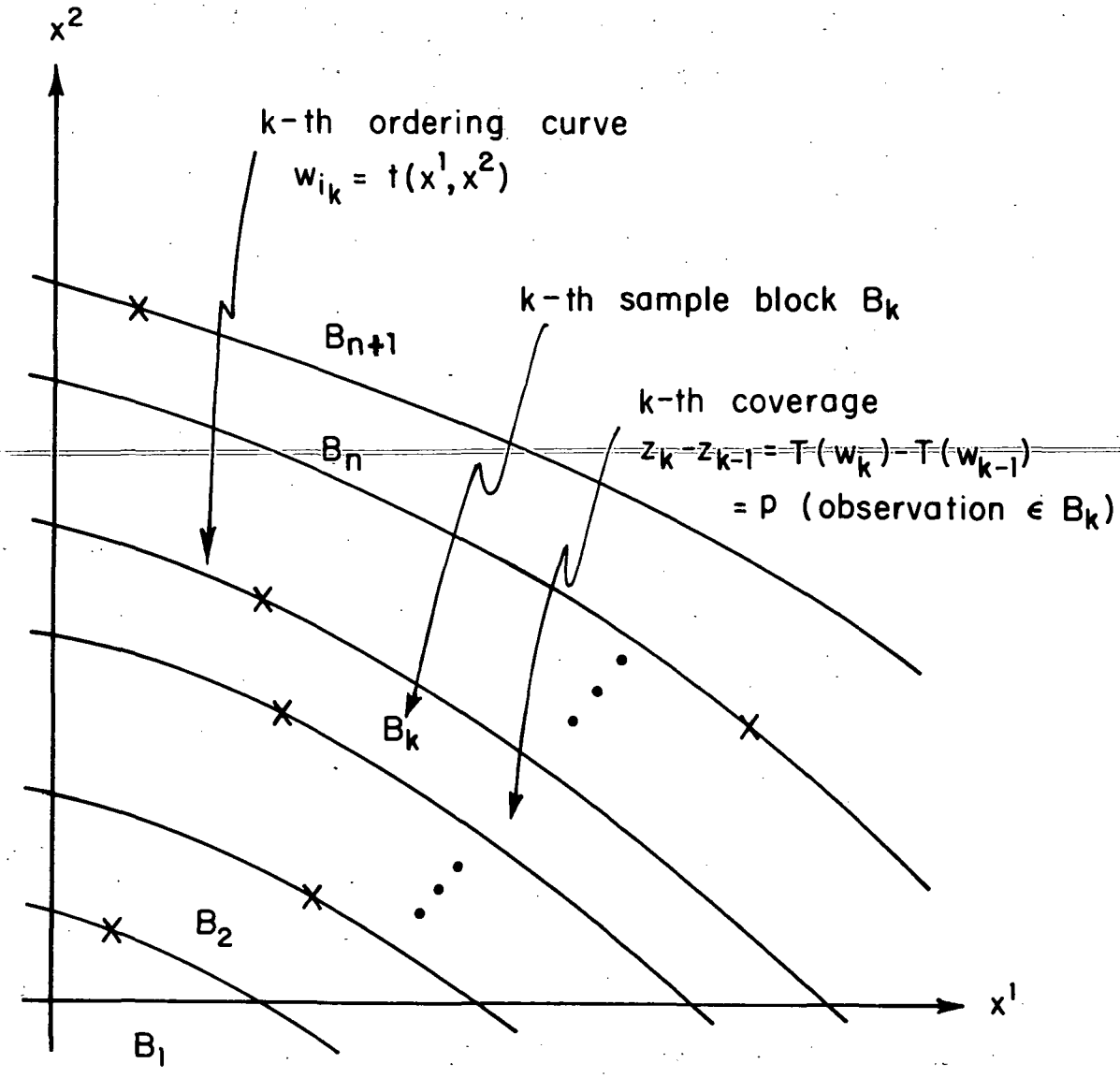


FIGURE II. 12

Example of Ordering Function

For a random sample of n random variables, it is also possible to partition the sample space into n+1 sample blocks by using as many as n different ordering functions. It can be shown [7], [10] that the coverages of each sample block, which are the probabilities of a future observation falling in each sample block, still follow the Beta distribution. Thus the expected value of the probability of a future observation falling in any r, $r \leq n$, of the sample blocks is $r/(n+1)$.

CHAPTER III

A SURVEY OF DENSITY FUNCTION ESTIMATES

Section I.6 of the introductory chapter mentioned that a classification method will be presented that uses density estimates in a sequential test called the sequential probability ratio test (SPRT). The chapters that follow this one examine density function estimates that are well suited for the SPRT and formulate an estimated version of the SPRT from the density estimates. Before proceeding to such a development, this chapter presents a survey of several known techniques for estimating density functions.

III.1 Assumptions

In discussing the density estimates presented in this report, the following assumptions about the samples from each class are made:

i) that the samples are scalars

ii) that the samples are independently, identically distributed in each class

iii) that the samples of each class are of the continuous type. (the footnote in Section II.2 defines a random variable of the continuous type.)

III.2 Motivation for Density Function Estimates

In order to get a clearer idea of what is involved in estimating a density function, the definition of a density function will be reviewed.

The probability distribution function F(x) of a random variable x is defined as $F(x) \triangleq p(\eta \leq x)$ and the density function f(x) is the derivative of F(x), $f(x) \triangleq \frac{dF(x)}{dx}$. In the pattern classification procedures discussed in this report, F(x) is unknown. The distribution function F(x) can be easily estimated from training samples by taking as the estimate the fraction of samples less than x (remember that only scalar samples are being treated in this chapter.) As the number of training samples approaches infinity, this estimate of F(x) approaches the true F(x) with probability one and in the mean square. Cramer [6] and Rao [13] are among many authors who discuss this estimate.

While the estimate of F(x) is straight forward, it is the estimate of f(x) = F'(x) that is actually needed. The definition of a derivative,

$$\lim_{h \to 0} \frac{F(x+h)-F(x-h)}{2h} = f(x) , \qquad \text{(III.1)}$$

can be used to motivate methods for estimating f(x). Equation (III.1) can be written more generally in terms of probabilities as

$$\lim_{\Delta \to 0} \frac{p(\text{observation } \varepsilon\ \Delta)}{\Delta} = f(x) \qquad \text{(III.2)}$$

where Δ is the width of some interval that contains x. Thus f(x) could be estimated by first approximating f(x) as in the left hand side of equation (III.1) or (III.2) and then estimating the approximation from training samples. Most methods which have been developed for estimating f(x) involve using equations (III.1) and (III.2) in one of two ways:

i) one approach is to specify the internal width Δ and and let the numerator p(observation ϵ Δ) be a random variable to be estimated from the training samples

ii) another approach is to specify the numerator p(observation ϵ Δ) and to specify a certain number of training samples to be contained in the interval Δ so that the denominator takes the value of that interval width Δ which contains the specified number of training samples.

In i) the interval width is specified and in ii) the training samples determine the interval width. Rosenblatt [14], Whittle [15], and Parzen [16] have written about i) and Loftsgaarden and Quesenberry [17] about ii). Cover [18] in a general discussion of nonparametric pattern recognition methods briefly discusses the use of the Parzen density estimate in a Bayes decision rule and mentions the estimate of Loftsgaarden and Quesenberry. The remainder of this chapter will discuss several density estimates stressing properties which are important to sequential decision methods where, of course, a string of observations are considered at once. Some considerations to be described are storage requirements, complexity of calculations, and continuity of the density estimates.

III.3 Density Models That Specify Bin Width

III.3.1 Fixed Bin Model

Perhaps the simplest density function estimate is the estimate that is often referred to as a histogram and what will be called the fixed bin model in this report. Referring to equation (III.2), this

density model sets the denominator and estimates the numerator. The sample axis is partitioned into a number of fixed intervals as in Figure III.1. The density estimate for an x in any interval is the fraction of training samples in that interval divided by the interval width. Let

n be the number of training samples

k be the number of bins

γ_i, $i=1,2,\ldots,k+1$ be the bin boundaries

m_i be the number of samples in the i-th bin (or in interval (γ_i,γ_{i+1})),

then

$$\hat{f}(x) = \begin{cases} \frac{m_i}{n}(\gamma_{i+1}-\gamma_i) & \text{for } \gamma_i < x < \gamma_{i+1} \\ 0 & \text{for } x < \gamma_1 \text{ or } x > \gamma_{k+1}. \end{cases} \qquad \text{(III.3)}$$

By its construction, estimate (III.3) is a step function. Since the intervals are specified by the choice of the γ_i's, only the γ_i's and the fraction of samples in each bin need be stored while using the estimate. Thus, the estimate is calculated for all x at once, and the whole density estimate is stored for future use. One question that must be answered in formulating this estimate is that of where to place the bins along the sample axis. If the bins are wide or are placed where there are few samples, the estimate $\hat{f}(x)$ may be inaccurate, and poor use will have been made of the training samples.

Hughes [19] discusses the effect of the number of training samples and the number of bins on the mean accuracy of a Bayes decision rule which uses

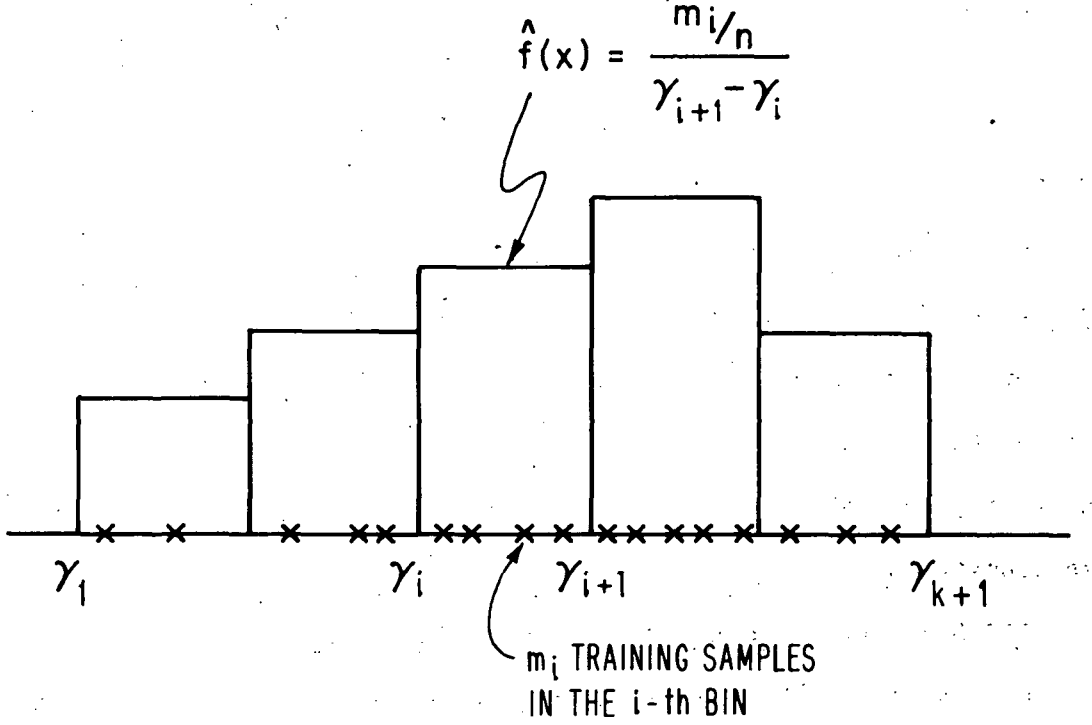


FIG. III.1

Example of Fixed Bin Density Estimate

fixed bin density estimates. In Hughes' paper, the placement of the training samples in the bins was given a uniform prior distribution in order to consider all possible combinations in which the training samples might occur in the bins. Abend and Harley [20], Chandrasekaran and Harley [21], and Hughes [22] amend the results of this paper by using the training samples to provide posterior estimates of the probabilities of an observation following in each bin so that the estimates will be consistant with the uniform prior distribution. Patrick and Hancock [23] examine the Bayes decision rule for problems where the training samples are available but their classification is unknown. In discussing the situation when no information is known about the density functions, they show that a fixed bin model can still be used to estimate the density functions.

III.3.2 Parzen Model (Specified Sliding Bin)

Parzen [16] estimates the density function at x by centering a bin of specified width about x. Similar to the fixed bin model, Parzen's density estimate specifies the denominator of equation (III.2) and estimates the numerator. As the bin (x-h,x+h) is always centered at the x for which the density estimate is desired, the mechanism of the model may be viewed as a sliding window of width 2h. Figure III.2 illustrates the model. The estimate at any x is

$$\hat{f}(x) = \frac{\text{fraction of training samples in } (x-h,x+h)}{2h} \qquad \text{(III.4)}$$

The model is similar to the fixed bin model in that the bin width is

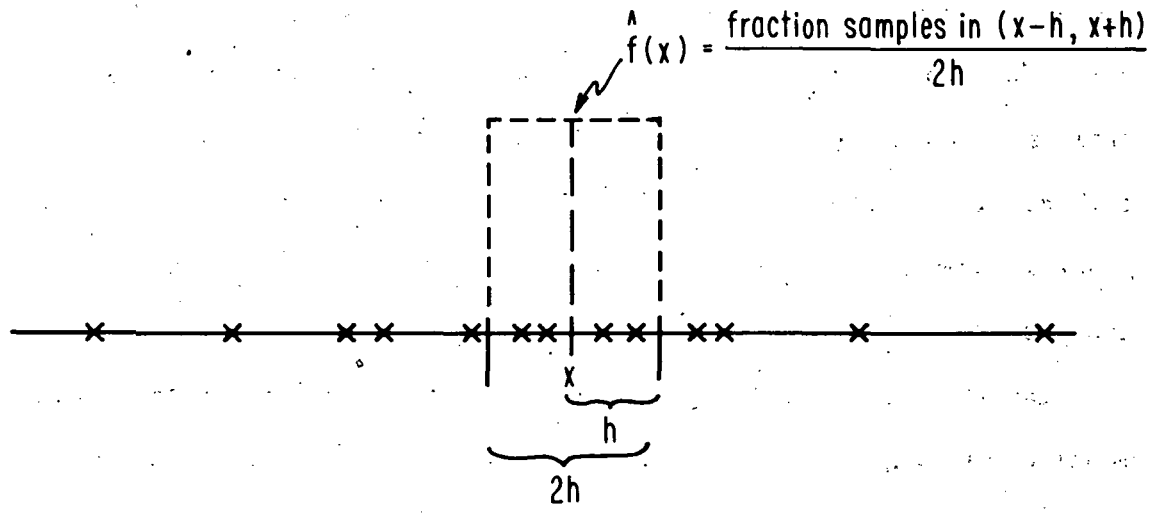


FIG. III.2

Example of Parzen Density Estimate

specified, and again there is the question of how wide to set the bin. It may be that for some x the interval (x-h,x+h) does not contain a great enough percentage of the training samples to provide an accurate estimate of f(x). Given an x, it may be necessary to change h until a satisfactory number of samples is contained in (x-h,x+h). Parzen and Rosenblatt have developed formulas for h as a function of the number of samples so that h minimizes the mean square error of the estimate, but these expressions require a knowledge of f(x) and usually f"(x). The utilization of this model in a decision algorithm requires that all training samples must be stored. The estimate is then calculated for each x. The estimate in equation (III.4) is not continuous, but the general formula for the Parzen estimator presented in the next paragraph can provide a continuous estimate.

Let there be n training samples $\{x_i\}$, i=1,2,...,n. Then Parzen's model can be expressed in a general formula

$$\hat{f}(x) = \frac{1}{nh(n)} \sum_{i=1}^{n} K\left(\frac{x-x_i}{h(n)}\right) \qquad \text{(III.5)}$$

where

$$\sup_{-\infty<y<\infty} |K(y)| < \infty$$

$$\int_{-\infty}^{\infty} |K(y)|dy < \infty$$

$$\lim_{y\to\infty} |yK(y)| = 0$$

$$\int_{-\infty}^{\infty} K(y)dy = 1$$

are conditions necessary for equation (III.5) to asymptotically be an unbiased estimator of f(x). The estimate (III.5) converges to f(x) in the mean square if $h(n) \to 0$ and $nh(n) \to \infty$ as $n \to \infty$. The convergence condition $h(n) \to 0$ may be interpreted in equation (III.4) as letting the interval width shrink to zero while the condition $nh(n) \to \infty$ requires the number of samples in the interval to approach infinity.

If

$$K(y) = \begin{cases} \frac{1}{2} & \text{for } |y| \leq 1 \\ 0 & \text{for } |y| > 1 \end{cases} \tag{III.6}$$

then equation (III.5) agrees with equation (III.4). The Parzen estimate is continuous in x for other choices of K(y). An example of K(y) which results in a continuous estimate is to take

$$K(y) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}y^2} \tag{III.7}$$

For this choice of K(y), estimate (III.5) is the sum of n Gaussian densities when each Gaussian density is centered about a training sample.

Van Ryzin [24] has developed a classification procedure that makes use of the Parzen estimator in Bayes rule.

III.4 Density Models where the Bin Width is Determined by Training Samples

III.4.1 Nearest Neighbor Density Estimate (Variable Sliding Bin)

Loftsgaarden and Quesenberry [17] have developed an estimate that employs an interval which is centered at x and whose width is determined

by the training samples. Unlike the fixed bin and Parzen models, the estimate of Loftsgaarden and Quesenberry specifies the numerator of equation (III.2) and estimates the denominator. In Section III.3.2, it was mentioned that the Parzen model could be viewed as a sliding bin of specified width centered at x. Similarly the Loftsgaarden and Quesenberry estimate can be viewed as a sliding bin of variable width. An integer $\ell(n)$ is chosen (n is always taken to be the number of training samples), and the $\ell(n)$-th nearest training sample to x, called $x_{\ell(n)}$ is found. The interval width is then taken to be $2|x-x_{\ell(n)}|$, and it follows that the fraction of samples inside the interval is $(\ell(n)-1)/n$. The estimate is

$$\hat{f}(x) = \frac{\ell(n) - 1}{n} \Big/ 2|x-x_{\ell(n)}| \qquad \text{(III.8)}$$

where $x_{\ell(n)}$ is the $\ell(n)$-th nearest sample to x according to the distance measure $|x-y|$. Figure III.3 provides an example. The estimate (III.8) converges to f(x) in probability if $\ell(n) \to \infty$ and $\ell(n)/n \to 0$ as $n \to \infty$. The condition $\ell(n)/n \to 0$ lets the width $w|x-x_{\ell}(n)|$ shrink to zero while the condition $\ell(n) \to \infty$ allows the number of training samples contained in the interval to approach infinity.

Metrics other than $|x-y|$ may be used in the estimate. In general, if the metric d(x,y) is employed, the estimate is

$$\hat{f}(x) = \frac{\ell(n)-1}{n} \Big/ 2d(x,x_{\ell(n)}) \qquad \text{(III.9)}$$

where $x_{\ell(n)}$ is the $\ell(n)$-th closest training sample to x according to the metric d(x,y).

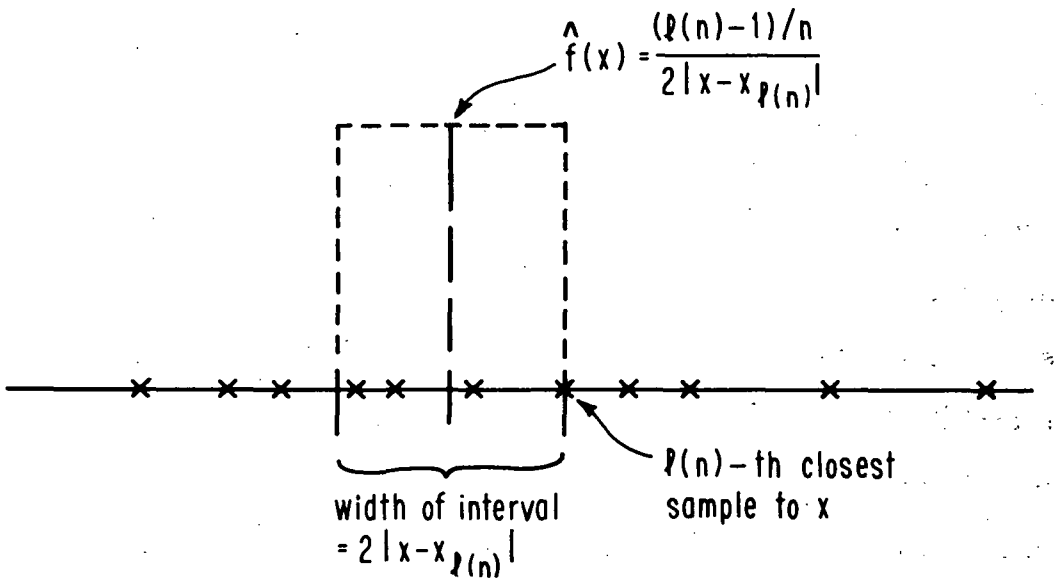


FIG. III.3

Example of Nearest Neighbor Density Estimate

The estimate of Loftsgaarden and Quesenberry is related to the nearest neighbor methods of pattern recognition [25, 26]. In the nearest neighbor (NN) methods, an observation is classified into that class which is most heavily represented among some specified number of nearest neighbors of the observation. Since the estimate of Loftsgaarden and Quesenberry involves finding the $\ell(n)$-th nearest neighbor to x, it will be called the nearest neighbor (NN) density estimate in this thesis.

The NN density estimate is continuous in x. All training samples must be stored in order to use the estimate, and then for any particular sample value x, the estimate is calculated. In the NN estimate, the bin centered at any x always contains a specified number of training samples; whereas in the Parzen estimate, in which the bin width is specified before hand, the interval may contain so few samples that the estimate can be quite inaccurate. This problem of bin placement is discussed further in Sections III.5 and IV.3.2.

III.5 Accuracy and Storage of Density Estimates

The purpose of studying density function estimates in this report is to examine their use in sequential classification algorithms. In practical decision problems, the amount of storage available for storing the density estimates during computation is limited. While limiting the storage of the density estimate is necessary, the accuracy of the estimate is thereby decreased.

In considering the accuracy of estimates of continuous density functions, the accuracy may be divided into two parts, one of a

deterministic nature and the other of a random nature. Density estimates make a deterministic approximation of f(x) in the neighborhood of x and then estimate the value of the approximation from the training samples. Thus, the training samples are not used to estimate f(x) directly but rather to estimate some deterministic approximation to f(x), which is a function of F(x), such as

$$\frac{F(x+h) - F(x-h)}{2h} \quad . \tag{III.10}$$

The total accuracy of the estimated density depends on how accurate an estimate of the approximation can be obtained from the training samples (the random part) and on the accuracy of the approximation (the deterministic part.)

For example in the Parzen estimate of equation (III.4), the density function is approximated by $[F(x+h) - F(x-h)]/2h$. The interval width $2h$ is specified, and then $F(x+h) - F(x-h)$ is estimated from the training samples. No matter how accurately $F(x+h) - F(x-h)$ is estimated, the accuracy of the Parzen estimate will be low if $[F(x+h) - F(x-h)]/2h$ is a poor approximation of f(x). Likewise, if $F(x+h) - F(x-h)$ is poorly estimated, the density estimator will be inaccurate even though $[F(x+h) - F(x-h)]/2h$ may accurately approximate f(x). Both the deterministic and random parts of a density estimate must be good for the total estimate to be accurate. The conditions for convergence of equation (III.5) express this phenomenon. The condition $h(n) \to 0$ requires the the interval width to shrink to zero and thus the deterministic part to converge; $n \to \infty$ causes the estimate of $F(x + h) - F(x - h)$ and hence the random part to converge. Both the

random and deterministic parts must converge simultaneously. The condition $nh(n) \to \infty$ means that as the interval width shrinks to zero the number of samples inside the interval approaches infinity. In general, the deterministic part of the accuracy depends on the bin size and the random part on the number of training samples including the number of samples inside the interval. The choice of the bin size is a trade off between making it small to provide deterministic accuracy or large to give random accuracy by containing a large fraction of training samples. Rosenblatt [14] shows that density estimates must be biased for a finite number of samples. The bias arises from the deterministic approximation of $f(x)$. The estimate of the approximation can be unbiased, but the error in the approximation still remains.

Since the intervals of the Parzen and NN estimates are centered at x, they are more accurate in the deterministic sense than the fixed bin model. But the Parzen and NN methods require storage of all training samples for good random accuracy. The fixed bin model sacrifices some deterministic accuracy but retains good random accuracy in limited storage.

This chapter has discussed some properties of different density estimates, but a more detailed discussion will be presented in the next chapter in connection with a new proposed estimate. The various properties of the density estimates discussed so far seem to be determined by two factors, 1.) whether the bin width is specified or is set by the training samples and 2.) whether the density function is estimated for all x at once and the total estimate stored, or all the training samples

are stored and the density is estimated separately for each x. Table III.1 lists the density estimates in a matrix form and shows how the various estimates are related to these two factors. Also listed are properties of the density estimates as determined by the two factors.

There is one blank position in the two by two matrix in Table III.1, and the next chapter will propose a density estimate that fits into the blank slot. The density estimate will combine some properties of the NN and fixed bin estimates as the blank position in the matrix indicates it should. The model will be a step function so the small storage advantage of the fixed bin model will be retained. But the bin widths and positions will be determined by the training samples so that the bin placement will result in an accuracy greater than the fixed bin model.

		Factor 2		Properties Influenced by Factor 1			
				In $f(x) \simeq p(x \epsilon \Delta)/\Delta$			
		Total Point Estimate[1]	Single Point Estimate[2]	denominator specified, numerator estimated	numerator specified, denominator estimated	Difficulty of bin size choice[4] more / less	Convergence conditions as # training samples→∞
Factor 1	Bin Width Set by Training Samples		NN 17,18[3]		✓	✓	specified bin width→0 at such a rate that # samples in bin→∞
	Bin Width Specified	Fixed Bin 19,20 21,22,23	Parzen 16,18 24	✓		✓	# samples specified in bin→∞ at such a rate that bin width→0[5]
Properties Influenced by Factor 2							
Is bin centered at x?	no	✓					
	yes		✓				
Storage requirement	small	✓					
	large		✓				
Computational complexity for any x	less	✓					
	more		✓				
Accuracy in deterministic sense	less	✓					
	more		✓				

1. In Total Point Estimate, the density function is estimated for all x at once, and the total estimate is stored.

2. In Single Point Estimate, all training samples are stored and the density is estimated separately for each x.

3. These numbers indicate references in the bibliography.

4. When the bin width is specified, there is a problem of how to choose it initially so as to contain a number of training samples that would give a reasonable estimate. In letting the training samples set the bin width, a reasonable estimate is more readily obtained.

5. The number of samples specified in the bin→∞ but a rate sufficiently slower than the total number of training samples→∞ in order that the bin width that contains the specified number of samples→0.

TABLE III.1 Properties of Fixed Bin, Parzen, and NN Density Estimates

CHAPTER IV

RANDOM BIN MODEL

Chapter III discussed three density estimates: the fixed bin model, the Parzen model, and the nearest neighbor model. Both the fixed bin and Parzen models have a computational disadvantage in that the bin width is specified before the density is estimated from the training samples. It is not known where to position the intervals in relation to the distribution of the training samples, and it is possible that the bin width could be set so wide as to contain half or even all of the training samples. If an interval contains too large a percentage of samples, the bin width can be changed and the density estimate repeated. But iterating on the interval width complicates the estimation of the density. The NN estimate overcomes the problem of setting the bin size by determining the interval width from the training samples. The number of training samples ℓ to be contained in a bin is specified, and the bin size is determined by the width necessary to contain this number of samples. Different values of ℓ result in different estimate accuracies, but whatever percentage of samples for a bin is specified, the bin width will be reasonable since it is determined by the distribution of the training samples. The density estimate presented in this chapter combines the property of the NN estimate of placing the bins by the training samples with the low storage advantage of the fixed bin model. Since the new density estimate has a step function form similar to the fixed bin model and at the same time determines the bin widths from

the training samples, the estimate is called the random bin density estimate.

IV.1 Presentation of Random Bin Estimate

The random bin model attempts to place bins so that the probability of an observation falling in each bin has a specified value. Usually the bins are positioned so it is equally likely an observation will fall in any bin is illustrated in Figure IV.1. Let k+1 be the number of bins. The bin widths are determined so the probability of an observation falling in any bin is $\frac{1}{k+1}$. Then

$$\hat{f}(x) = \frac{1}{k+1} \Big/ \left[\begin{array}{l}\text{estimated width of i-th} \\ \text{bin such that } p(x \varepsilon \text{i-th bin}) = \frac{1}{k+1}\end{array}\right] \quad \text{for } x \varepsilon \text{i-th bin} \tag{IV.1}$$

The bin boundaries are calculated from quantiles and quantile estimates. The next few sections discuss quantiles, their estimates, and a density estimate based on quantiles. The assumptions on the data listed in Section III.1 still hold in the following discussion. The assumptions were that the samples are scalars, identically and independently distributed in each class with absolutely continuous distribution functions. Conditions for the density estimates discussed in this thesis to converge to the true density f(x) require that f(x) be continuous at x. By the assumption of absolute continuity of F(x), the number of discontinuous points of f(x) is finite in any finite interval. Since in this report the purpose of obtaining density estimates is to classify observations, a density is estimated

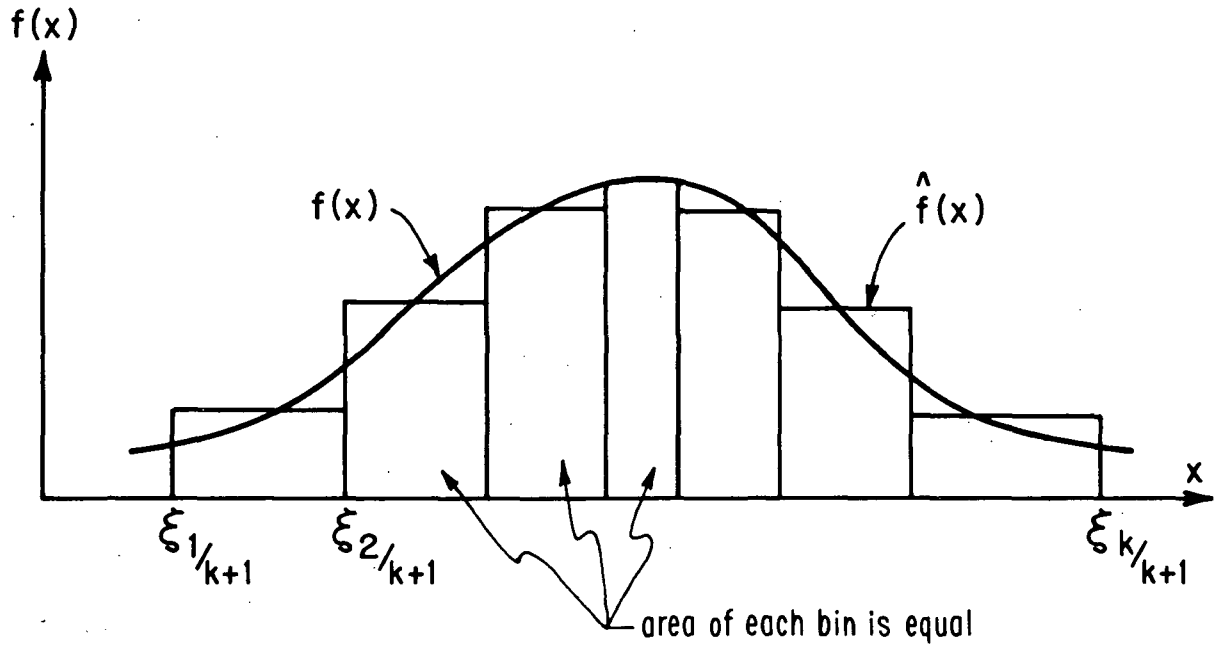


FIG. IV.1

Bin Placement for Random Bin Density Estimate

only at values of a given observation. The probability of an observation occuring at a discontinuity is zero. Thus, the assumption of absolute continuity of F(x) is not restrictive for classification purposes. The convergence conditions of the random bin density estimate that will be presented in Theorem IV.2 also assume f'(x) is continuous in a neighborhood of x and $f(x) \neq 0$ at x. Again, as long as the number of points at which f(x) is not continuously differentiable or f(x) equal to zero is finite in any finite interval, the conditions are not restrictive.

IV.1.1 Definition of Quantile

The p-th order quantile, labeled ξ_p, of a distribution function F(x) is any value of x such that $F(x=\xi_p) = p$. See Figure IV.2. In this report, ξ_p is assumed to be unique for any p. Since x is a random variable of the continuous type and hence F(x) is absolutely continuous, the existence of ξ_p for any p is guaranteed. The further assumption of the uniqueness of ξ_p means that F(x) is strictly increasing in x.

IV.1.2 Set of Quantiles

For any integer k, a set of k quantiles $(\xi_{\frac{1}{k+1}}, \xi_{\frac{2}{k+1}}, \ldots, \xi_{\frac{k}{k+1}})$ can be defined such that for any two consecutive quantiles $\xi_{\frac{j}{k+1}}$ and $\xi_{\frac{j+i}{k+1}}$,

$$F(\xi_{\frac{j+1}{k+1}}) - F(\xi_{\frac{j}{k+1}}) = \frac{1}{k+1} \quad . \qquad \text{(IV.2)}$$

Figure IV.3 provides an illustration. Thus the set of k quantiles

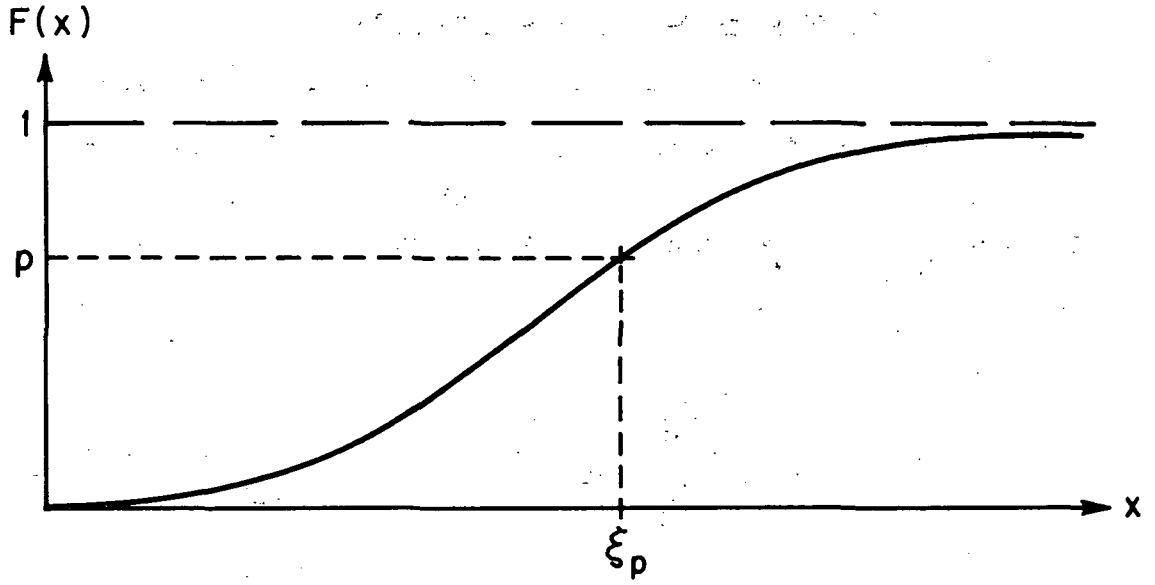


FIG. IV.2

Example of p-th Order Quantile

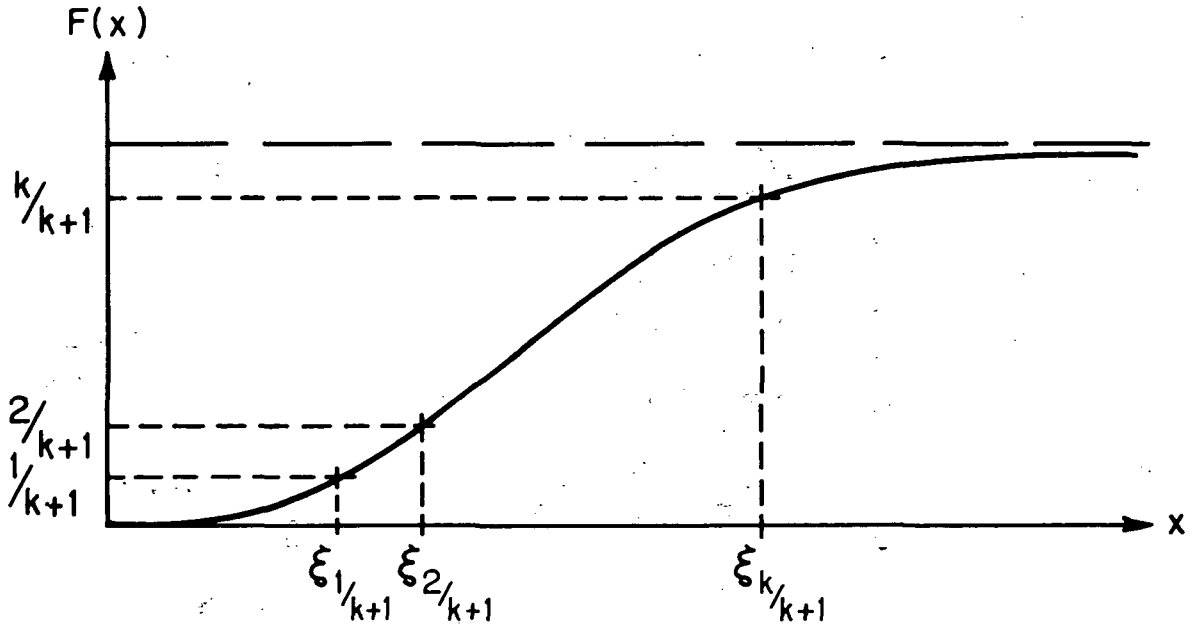


FIG. IV.3

Example of Set of Quantiles

partition the sample axis so that the probability of an observation falling in any partition is $\frac{1}{k+1}$.

IV.1.3 Defining a Density from Quantiles

Let the X be a random variable with distribution function F(x) and density f(x), and let $\xi_{\frac{1}{k+1}}$, $\xi_{\frac{2}{k+1}}$, ... , $\xi_{\frac{k}{k+1}}$ be a set of k quantiles. An approximation of f(x) is

$$f_{approx}(x) = \begin{cases} 0 & x < \xi_{\frac{1}{k+1}} \\ \dfrac{F(\xi_{\frac{j+1}{k+1}}) - F(\xi_{\frac{j}{k+1}})}{\xi_{\frac{j+1}{k+1}} - \xi_{\frac{j}{k+1}}} & \xi_{\frac{j}{k+1}} \leq x \leq \xi_{\frac{j+1}{k+1}} \\ 0 & x > \xi_{\frac{k}{k+1}} \end{cases} \qquad \text{(IV.3)}$$

If X is known to be distributed over an interval (a,b) then $f_{approx}(x)$ can be written as

$$f_{approx}(x) = \begin{cases} 0 & x \leq a \\ F(\xi_{\frac{1}{k+1}}) \Big/ (\xi_{\frac{1}{k+1}} - a) & a \leq x \leq \xi_{\frac{1}{k+1}} \\ \dfrac{F(\xi_{\frac{j+1}{k+1}}) - F(\xi_{\frac{j}{k+1}})}{\xi_{\frac{j+1}{k+1}} - \xi_{\frac{j}{k+1}}} & \xi_{\frac{j}{k+1}} \leq x \leq \xi_{\frac{j+1}{k+1}} \\ [1 - F(\xi_{\frac{k}{k+1}})] \Big/ (b - \xi_{\frac{k}{k+1}}) & \xi_{\frac{k}{k+1}} \leq x \leq b \\ 0 & x > b \end{cases}$$

By equation (IV.2), the numerators of equation (IV.3) are all equal to $\frac{1}{k+1}$.

If k is allowed to approach infinity and for any x one chooses from the set of k quantiles ($\xi_{\frac{1}{k+1}}, \xi_{\frac{2}{k+2}}, \ldots, \xi_{\frac{k}{k+1}}$) the pair of quantiles just below and just above x, the approximation converges to f(x). This is shown below:

Theorem 1: Let X be a random variable with an absolutely continuous function F(x) and with probability density function f(x). Let ($\xi_{\frac{1}{k+1}}, \xi_{\frac{2}{k+1}}, \ldots, \xi_{\frac{k}{k+1}}$) be a set of k quantiles from F(x). Define

$$f_{approx}(x) = \begin{cases} 0 & x < \xi_{\frac{1}{k+1}} \\ \dfrac{1/(k+1)}{\xi_{\frac{j+1}{k+1}} - \xi_{\frac{j}{k+1}}} & \xi_{\frac{j}{k+1}} \leq x \leq \xi_{\frac{j+1}{k+1}} \\ 0 & x > \xi_{\frac{k}{k+1}} \end{cases} \qquad \text{(IV.4)}$$

Then at all x for which f(x) is continuous

$$\lim_{k \to \infty} f_{approx}(x) = f(x) \qquad \text{(IV.5)}$$

First convergence of a more general form of equation (IV.4) will be proved.

Lemma 1: Let F(x) be an absolutely continuous function. For a constant x, let a_k be a sequence of real numbers such that $a_k \leq x$ and $a_k \to x$ as $k \to \infty$, and b_k be a sequence of real numbers such that

$b_k \geq x$ and $b_k \to x$ as $k \to \infty$. Then at all x for which F'(x) is continuous

$$\lim_{k\to\infty} \frac{F(b_k) - F(a_k)}{b_k - a_k} = F'(x) . \tag{IV.6}$$

Proof: Since F(x) is absolutely continuous,

$$\Delta F_k \triangleq \frac{F(b_k)-F(a_k)}{b_k-a_k} = \frac{1}{b_k-a_k} \int_{a_k}^{b_k} F'(\mu)d\mu . \tag{IV.7}$$

Subtracting F'(x) from both sides of equation (IV.7),

$$\Delta F_k - F'(x) = \frac{1}{b_k-a_k} \int_{a_k}^{b_k} (F'(\mu)-F'(x))d\mu . \tag{IV.8}$$

By the assumption of continuity of F(x) near x, for all $\varepsilon > 0$ there exists a $\delta_\varepsilon > 0$ such that $|F'(y)-F'(t)| < \varepsilon$ if $|y-t| < \delta_\varepsilon$. Given an ε, choose k_ε such that $b_k-a_k < \delta_\varepsilon$ if $k \geq k_\varepsilon$. Then it is observed that $|\mu-x|<\delta_\varepsilon$ if $k \geq k_\varepsilon$ and $a_k \leq \mu \leq b_k$ (remember $a_k \leq x \leq b_k$). The condition $|\mu-x|<\delta_\varepsilon$ and continuity of F'(x) imply

$$|F'(\mu) - F'(x)| < \varepsilon . \tag{IV.9}$$

Substituting equation (IV.9) in equation (IV.8),

$$|\Delta F_k - F'(x)| < \frac{1}{b_k-a_k} \int_{a_k}^{b_k} \varepsilon d\mu = \varepsilon \tag{IV.10}$$

and so $|\Delta F_k - F'(x)| < \varepsilon$ if $k \geq k_\varepsilon$. Thus for any $\varepsilon > 0$, there exists a k_ε such that $|\Delta F_k - F(x)| < \varepsilon$ if $k \geq k_\varepsilon$, and Lemma 1 is proven.

Proof of Theorem 1: Lemma 1 implies Theorem 1 if $\xi_{\frac{j}{k+1}}$ in equation (IV.4) can be identified with a_k and $\xi_{\frac{j+1}{k+1}}$ with b_k. By construction of equation (IV.4), $\xi_{\frac{j}{k+1}} \leq x \leq \xi_{\frac{j+1}{k+1}}$. It remains to show that $\xi_{\frac{j}{k+1}} \to x$ and $\xi_{\frac{j+1}{k+1}} \to x$ as $k \to \infty$. It has been assumed that for any p the p-th order quantile ξ_p is unique, and hence F(x) is strictly increasing in x. So $\xi_{\frac{j}{k+1}} \leq x \leq \xi_{\frac{j+1}{k+1}}$ implies

$$F(\xi_{\frac{j}{k+1}}) \leq F(x) \leq F(\xi_{\frac{j+1}{k+1}}) \quad . \tag{IV.11}$$

Now $F(\xi_{\frac{j+1}{k+1}}) - F(\xi_{\frac{j}{k+1}}) = \frac{1}{k+1}$. For any $\varepsilon > 0$, there exists a k_ε such that $\frac{1}{k+1} < \varepsilon$ if $k > k_\varepsilon$. Thus

$$\lim_{h \to \infty} [F(\xi_{\frac{j+1}{k+1}}) - F(\xi_{\frac{j}{k+1}})] = 0 \quad . \tag{IV.12}$$

Equations (IV.11) and (IV.12) imply

$$\lim_{k \to \infty} F(\xi_{\frac{j}{k+1}}) = F(x) \tag{IV.13}$$

and

$$\lim_{k \to \infty} F(\xi_{\frac{j+1}{k+1}}) = F(x) \quad . \tag{IV.14}$$

Since F(x) is strictly increasing in x, equations (IV.13) and (IV.14)

imply

$$\lim_{k\to\infty} \xi_{\frac{j}{k+1}} = x \qquad \text{(IV.15)}$$

and

$$\lim_{k\to\infty} \xi_{\frac{j+1}{k+1}} = x \quad . \qquad \text{(IV.16)}$$

The sequence $\xi_{\frac{j}{k+1}}$ has the properties of a_k and $\xi_{\frac{j+1}{k+1}}$ those of b_k, and so Lemma 1 implies Theorem 1.

IV.1.4 Quantile Estimates

Equation (IV.4) presents a density approximation containing quantiles. If F(x) is unknown, the quantile can be estimated from training samples. A density estimate can be constructed by replacing the quantiles in equation (IV.4) with quantile estimates.

The p-th order quantile of a distribution function F(x) can be estimated from training samples with order statistic theory. Let n independent observations of a random variable X be arranged in ascending order.

$$x_{i_1} < x_{i_2} < \ldots < x_{i_n} \quad . \qquad \text{(IV.17)}$$

Relabeled the samples for convenience

$$y_1 = x_{i_1}, \quad y_2 = x_{i_2}, \quad \ldots \quad , y_n = x_{i_n} . \qquad \text{(IV.18)}$$

$(y_1, y_2, \ldots, y_n)$ is, as is mentioned in Section II.3, a set of order statistics. An estimate of the p-th order quantile ξ_p is

$$\hat{\xi}_p = y_{[np]+1} \tag{IV.19}$$

where [w] is the largest integer less than or equal to w. If np is an integer, choose any value in the closed interval between y_{np} and y_{np+1} since the distance between the two neighboring order statistics y_{np} and y_{np+1} tends to zero as n approaches infinity. A motivation for $\hat{\xi}_p$ is that the fraction of samples less than $\hat{\xi}_p$ is near p and from order statistic theory (see Appendix II.3) $E[F(\hat{\xi}_p)] = \frac{[np]+1}{n+1}$ which is approximately p. Rao [13] shows that the estimate $\hat{\xi}_p$ approaches ξ_p as $n \to \infty$ with probability one. The distribution of ξ_p is shown by David [12] to be asymptotically Gaussian with mean ξ_p and variance $\frac{p(1-p)}{n[f(\xi_p)]^2}$ where for np equal an integer $\hat{\xi}_p$ is taken to be y_{np} to simplify the indeterminate case.

The set of quantiles $(\xi_{\frac{1}{k+1}}, \xi_{\frac{2}{k+1}}, \ldots, \xi_{\frac{k}{k+1}})$ that appear in the density approximation of equation (IV.4) can be estimated from equation (IV.19)

$$\hat{\xi}_{\frac{j}{k+1}} = y_{[\frac{jn}{k+1}]+1} \quad \text{for } j=1,2,\ldots,k \; . \tag{IV.20}$$

IV.1.5 Estimating the Density Function f(x)

If a set of k quantile estimates $(\hat{\xi}_{\frac{1}{k+1}}, \hat{\xi}_{\frac{2}{k+1}}, \ldots, \hat{\xi}_{\frac{k}{k+1}})$ determined by equation (IV.20) serve as the bin boundaries in the random bin estimate, then each bin contains approximately the same number of training samples. The random bin density estimate is

$$\hat{f}(x) = \begin{cases} 0 & x < \hat{\xi}_{\frac{1}{k+1}} \\ \frac{1}{k+1} \Big/ (\hat{\xi}_{\frac{j+1}{k+1}} - \hat{\xi}_{\frac{j}{k+1}}) & \hat{\xi}_{\frac{j}{k+1}} < x < \hat{\xi}_{\frac{j+1}{k+1}} \\ 0 & x > \hat{\xi}_{\frac{k}{k+1}} \end{cases} \tag{IV.21}$$

The following theorem shows that the random bin density estimate converges in probability to the true density if $k \to \infty$ and $k/n \to 0$ as $n \to \infty$. For convergence of $\hat{f}(x)$, the bin width must approach zero yet contain an infinite number of training samples. The condition $k \to \infty$ lets the bin width tend to zero while $k/n \to 0$ allows the number of samples in each bin to approach infinity. The need for $k \to \infty$ and $k/n \to 0$ as $n \to \infty$ can also be seen by inspecting $(k+1)(\hat{\xi}_{\frac{j+1}{k+1}} - \hat{\xi}_{\frac{j}{k+1}})$. The conditions $k \to \infty$ and $n \to \infty$ are necessary for $\hat{\xi}_{\frac{j+1}{k+1}} - \hat{\xi}_{\frac{j}{k+1}} \to 0$. Since $\hat{\xi}_{\frac{j+1}{k+1}} - \hat{\xi}_{\frac{j}{k+1}}$ is multiplied by $k+1$ and $k+1 \to \infty$, an additional condition of $k/n \to \infty$ is needed in order that both $(k+1)$ and $(\hat{\xi}_{\frac{j+1}{k+1}} - \hat{\xi}_{\frac{j}{k+1}})$ converge at rates appropriate for $\frac{1}{(k+1)(\hat{\xi}_{\frac{j+1}{k+1}} - \hat{\xi}_{\frac{j}{k+1}})}$ to converge to $f(x)$.

In the proof of convergence of $\hat{f}(x)$ to $f(x)$ in the following theorem, a lemma is first developed that shows that $\hat{f}(x)$ follows

asymptotically for large n a Gaussian distribution. The lemma shows that since $\hat{\xi}_{\frac{j}{k+1}}$, j=1,2,...,k, are asymptotically jointly Gaussian the asymptotic distribution of $1/\hat{f}(x) = (k+1)(\hat{\xi}_{\frac{j+1}{k+1}} - \hat{\xi}_{\frac{j}{k+1}})$, which is a linear combination of two Gaussian random variables, is Gaussian. The asymptotic distribution of $\hat{f}(x)$ is then proved to be Gaussian. The proof of the theorem concludes by showing the convergence of $\hat{f}(x)$ to f(x).

Theorem 2: Let $X_1, X_2, \ldots, X_n$ be n independent random variables identically distributed as a random variable X with an absolutely continuous distribution function F(x) and with probability density function f(x). Let $(Y_1, Y_2, \ldots, Y_n)$ be the set of n order statistics for $(X_1, X_2, \ldots, X_n)$, and let $\hat{\xi}_{\frac{j}{k(n)+1}} = Y_{[\frac{jn}{k(n)+1}]+1}$, j=1,2,...,k(n), where k(n) is a sequence of positive integers such that $k(n) \to \infty$ and $k(n)/n \to 0$ as $n \to \infty$. Define

$$\hat{f}_n(x) = \begin{cases} 0 & x < \hat{\xi}_{\frac{1}{k(n)+1}} \\ \frac{1}{k(n)+1} \Big/ (\hat{\xi}_{\frac{j+1}{k(n)+1}} - \hat{\xi}_{\frac{j}{k(n)+t}}) & \hat{\xi}_{\frac{j}{k(n)+1}} \leq x \leq \hat{\xi}_{\frac{j+1}{k(n)+1}} \\ 0 & x > \hat{\xi}_{\frac{k(n)}{k(n)+1}} \end{cases} \tag{IV.22}$$

Then $\hat{f}_n(x)$ is a consistent* estimator of $f(x)$ at all x in the neighborhood of which $f(x)$ and $f'(x)$ are continuous and $f(x) \neq 0$.

Before the theorem is proven, the following lemma is developed.

Lemma 2: The density estimate $\hat{f}_n(x)$ defined in Theorem 2 follows asymptotically a Gaussian distribution.

Proof of Lemma 2: First, $1/\hat{f}_n(x)$ will be shown to be asymptotically Gaussian. If s quantiles $\xi_{p_1}, \xi_{p_2}, \ldots, \xi_{p_s}$ are estimated by equation (IV.19) and $f(x)$ is differentiable in the neighborhood of ξ_{p_i}, then the s quantile estimates $\hat{\xi}_{p_1}, \hat{\xi}_{p_2}, \ldots, \hat{\xi}_{p_s}$ follow asymptotically an s-variate Gaussian distribution [12] with means

$$E_G\hat{\xi}_{p_i} = \xi_{p_i} \quad , \tag{IV.24}$$

variances

$$\mathrm{var}_G(\hat{\xi}_{p_i}) = \frac{p_i(1-p_i)}{n[f(\xi_{p_i})]^2} \quad , \tag{IV.25}$$

and covariances

$$\mathrm{cov}_G(\hat{\xi}_{p_i},\hat{\xi}_{p_j}) = \frac{p_i(1-p_j)}{nf(\xi_{p_i})f(\xi_{p_j})} \quad , \quad i < j \ . \tag{IV.26}$$

---

*Let $X_1, X_2, \ldots, X_n$ be n independent random variables identically distributed as a random variable X with distribution function $F(x)$. $\hat{\theta}(X_1, X_2, \ldots, X_n)$ is a consistent estimator of θ if $\hat{\theta}(X_1, X_2, \ldots, X_n)$ converges to θ as $n \to \infty$. Convergence in this report is shown in probability.

A subscript G has been used to indicate that these are the means and variances of the asymptotic Gaussian distribution since the means and variances of the asymptotic distribution of a random variable are not necessarily equal to the limits of the actual means and variances of the variable. Letting s=2, $p_1 = \frac{j}{k(n)+1}$ and $p_2 = \frac{j+1}{k(n)+1}$, then $(k(n)+1)(\hat{\xi}_{\frac{j+1}{k(n)+1}} - \hat{\xi}_{\frac{j}{k(n)+1}})$ is a linear combination of two asymptotically Gaussian random variables and so is itself asymptotically Gaussian with mean

$$E_G(k(n)+1)(\hat{\xi}_{\frac{j+1}{k(n)+1}} - \hat{\xi}_{\frac{j}{k(n)+1}}) = (k(n)+1)(\xi_{\frac{j+1}{k(n)+1}} - \xi_{\frac{j}{k(n)+1}}) \qquad \text{(IV.27)}$$

and variance

$$\begin{aligned}
&\mathrm{var}_G[(k(n)+1)(\hat{\xi}_{\frac{j+1}{k(n)+1}} - \hat{\xi}_{\frac{j}{k(n)+1}})] \\
&= \frac{(k(n)+1)^2}{n}\left[\mathrm{var}_G(\hat{\xi}_{\frac{j+1}{k(n)+1}}) - 2\,\mathrm{cov}_G(\hat{\xi}_{\frac{j+1}{k(n)+1}}, \hat{\xi}_{\frac{j}{k(n)+1}})\right. \\
&\qquad \left. + \mathrm{var}_G(\hat{\xi}_{\frac{j}{k(n)+1}})\right] \\
&= \frac{(k(n)+1)^2}{n}\left\{\frac{\frac{j+1}{k(n)+1}(1-\frac{j+1}{k(n)+1})}{[f(\xi_{\frac{j+1}{k(n)+1}})]^2} - 2\frac{\frac{j}{k(n)+1}(1-\frac{j+1}{k(n)+1})}{f(\xi_{\frac{j}{k(n)+1}})f(\xi_{\frac{j+1}{k(n)+1}})}\right. \\
&\qquad \left. + \frac{\frac{j}{k(n)+1}(1-\frac{j}{k(n)+1})}{[f(\xi_{\frac{j}{k(n)+1}})]^2}\right\} . \qquad \text{(IV.28)}
\end{aligned}$$

These two equations are actually the asymptotic means and variances of $1/\hat{f}_n(x)$. Before finding the asymptotic mean and variance of $\hat{f}_n(x)$, it will be shown that $\text{var}_G\left(\frac{1}{\hat{f}_n(x)}\right)$ tends to 0 as $n \to \infty$.

This will be shown by expanding the terms in equation (IV.28).

By definition of quantiles, $\frac{j}{k(n)+1} = F(\xi_{\frac{j}{k(n)+1}})$ and $\frac{j+1}{k(n)+1} = F(\xi_{\frac{j+1}{k(n)+1}})$. For convenience, let $h_1 = x - \xi_{\frac{j}{k(n)+1}}$ and $h_2 = \xi_{\frac{j+1}{k(n)+1}} - x$. $F(x-h_1)$, $F(x+h_2)$, $f(x-h_1)$, and $f(x+h_2)$ can be expanded to

$$F(x-h_1) = F(x) - h_1 f(x) + \frac{h_1^2}{2} f'(\theta) \qquad \frac{\xi_j}{k(n)+1} < \theta < x \,, \tag{IV.29}$$

$$F(x+h_2) = F(x) + h_2 f(x) + \frac{h_2^2}{2} f'(\phi) \qquad x < \phi < \xi_{\frac{j+1}{k(n)+1}} \tag{IV.30}$$

$$\frac{1}{f(x-h_1)} = \frac{1}{f(x)} + h_1 \frac{f'(\gamma)}{[f(\gamma)]^2} \qquad \xi_{\frac{j}{k(n)+1}} < \gamma < x \tag{IV.31}$$

and

$$\frac{1}{f(x+h_2)} = \frac{1}{f(x)} - h_2 \frac{f'(\mu)}{[f(\mu)]^2} \qquad x < \mu < \xi_{\frac{j+1}{k(n)+1}} \tag{IV.32}$$

After substituting the above four equations into the expression for

$\text{var}_G(\frac{1}{\hat{f}(x)})$ in equation (IV.28) and performing algebraic manipulations, the $\text{var}_G(\frac{1}{\hat{f}(x)})$ becomes

$$\text{var}_G\left(\frac{1}{\hat{f}(x)}\right) = \frac{(k(n)+1)^2}{n}\left\{\frac{(h_1+h_2)}{f(x)} + 0(h_1^2) + 0(h_2^2) + 0(h_1h_2)\right\} . \tag{IV.33}$$

Now an expression for (k(n)+1) will be found. Upon subtracting equation (IV.29) from equation (IV.30),

$$F(x+h_2)-F(x-h_1) = f(x)(h_1+h_2) + 0(h_1^2) + 0(h_2^2) \tag{IV.34}$$

Since $F(x+h_2)-F(x-h_1) = 1/(k(n)+1)$, it is found after algebraic manipulations that

$$k(n)+1 = \frac{1}{h_1+h_2}\left\{\frac{1}{f(x)+[\theta(h_1^2)+0(h_2^2)]/(h_1+h_2)}\right\} . \tag{IV.35}$$

Substituting this expression into the $\text{var}_G[1/\hat{f}_n(x)]$ in equation (IV.33),

$$\text{var}_G[1/\hat{f}_n(x)] = \frac{k(n)+1}{n}\left\{\frac{1}{f(x)+[0(h_1^2)+0(h_2^2)]/(h_1+h_2)}\right\} \cdot\left\{\frac{1}{f(x)} + \frac{0(h_1^2)+0(h_2^2)+0(h_1h_2)}{h_1+h_2}\right\} . \tag{IV.36}$$

Equations (IV.15) and (IV.16) in the proof of Lemma IV.1 state that $\xi_{\frac{j}{k(n)+1}} \to x$ and $\xi_{\frac{j+1}{k(n)+1}} \to x$ as $k(n) \to \infty$, and so $h_1 \to 0$ and $h_2 \to 0$ as $k(n) \to \infty$. Since $k(n)/n \to 0$ as $n \to \infty$,

$$\lim_{n\to\infty} \operatorname{var}_G[1/\hat{f}_n(x)] = 0 \qquad \text{(IV.37)}$$

Now that the asymptotic distribution of $1/\hat{f}_n(x)$ has been found and it has been shown that $\operatorname{var}_G[1/\hat{f}_n(x)] \to 0$ as $n \to \infty$, the asymptotic distribution of $\hat{f}_n(x)$ will be obtained by the following:

Lemma (David [12]): Let $X_1, X_2, \ldots, X_n$ be n independent random variables identically distributed as a random variable X. Then $t_j(X_1, X_2, \ldots, X_n)$, $j=1,2,\ldots,m$, are m random variables that are functions of $(X_1, X_2, \ldots, X_n)$. If the random variables $t_j(X_1, X_2, \ldots, X_n)$, $j=1,2,\ldots,m$, have asymptotically an m-variate Gaussian distribution with means μ_j, variances σ_j^2 which tend to 0 as $n \to \infty$, and covariances σ_{ij}, and if $g_j(t_j)$ are single-valued functions with nonvanishing continuous derivatives $g_j'(t_j)$ in the neighborhoods of $t_j = \mu_j$, then $g_j(t_j)$ themselves have an m-variate Gaussian distribution with means $g_j(\mu_j)$ and covariance $\sigma_{ij} g_i'(\mu_i) g_j'(\mu_j)$.

With $m = 1$, $t = (k(n)+1)(\hat{\xi}_{\frac{j+1}{k(n)+1}} - \hat{\xi}_{\frac{j}{k(n)+1}})$, and $\mu = \frac{1}{f(x)} = (k(n)+1)(\xi_{\frac{j+1}{k(n)+1}} - \xi_{\frac{j}{k(n)+1}})$ and since $f(x) \neq 0$, the transformation $g(t) = \frac{1}{t}$ satisfies the conditions of the lemma, Since $g'(t) = \frac{-1}{t^2}$, $\hat{f}_n(x)$ is asymptotically Gaussian with mean

$$E_G \hat{f}_n(x) = \frac{1}{k(n)+1} \Big/ (\xi_{\frac{j+1}{k(n)+1}} - \xi_{\frac{j}{k(n)+1}}) \qquad \text{(IV.38)}$$

and variance

$$\operatorname{var}_G[\hat{f}_n(x)] = [\frac{1}{k(n)+1} \Big/ (\xi_{\frac{j+1}{k(n)+1}} - \xi_{\frac{j}{k(n)+1}})]^4 \operatorname{var}[1/\hat{f}_n(x)] \ . \qquad \text{(IV.39)}$$

Theorem 1 states that

$$\lim_{n\to\infty} \frac{1}{k(n)+1} \Big/ \left(\xi_{\frac{j+1}{k(n)+1}} - \xi_{\frac{j}{k(n)+1}}\right) = f(x) \ , \tag{IV.40}$$

and so

$$\lim_{n\to\infty} E_G \hat{f}_n(x) = f(x) \quad . \tag{IV.41}$$

Further, since $\mathrm{var}_G[1/\hat{f}(x)] \to 0$ as $n \to \infty$,

$$\lim_{n\to\infty} \mathrm{var}_G[\hat{f}_n(x)] = 0 \tag{IV.42}$$

Lemma 2 has been proven.

Proof of Theorem 2: From Lemma 2, $\hat{f}_n(x)$ follows asymptotically the Gaussian distribution $\Phi_n(u)$ with mean $E_G(\hat{f}_n(x))$ and variance $\mathrm{var}_G[\hat{f}(x)]$,

$$\Phi_n(u) = \int_{-\infty}^{\frac{u-E_G(\hat{f}_n(x))}{(\mathrm{var}_G[\hat{f}_n(x)])^{1/2}}} \frac{1}{\sqrt{2\pi}} e^{-1/2\, v^2} dv \ . \tag{IV.43}$$

From Lemma 2, $E_G(\hat{f}_n(x)) \to f(x)$ and $\mathrm{var}_G[\hat{f}_n(x)] \to 0$ as $n \to \infty$, so

$$\lim_{n\to\infty} \frac{u-E_G(\hat{f}_n(x))}{(\mathrm{var}_G[\hat{f}_n(x)])^{1/2}} = \begin{cases} -\infty & u < f(x) \\ 0 & u = f(x) \\ \infty & u > f(x) \end{cases} \tag{IV.44}$$

Thus

$$\lim_{n\to\infty} \Phi_n(u) = \begin{cases} 0 & u < f(x) \\ \frac{1}{2} & u = f(x) \\ 1 & u > f(x) \end{cases} \qquad \text{(IV.45)}$$

Define

$$F(u) = \begin{cases} 0 & u < f(x) \\ 1 & u \geq f(x) \end{cases} \qquad \text{(IV.46)}$$

The limit of $\Phi_n(u)$ as $n \to \infty$ equals $F(u)$ at all points for which $F(u)$ is continuous. Since $\Phi_n(u)$ is degenerate at $u = f(x)$ in the limit, $\hat{f}_n(x)$ converges in probability to $f(x)$.

IV.2 Restatement of Algorithm for Random Bin Density Estimate

This section presents a concise summary of the procedure for finding the random bin density estimate from n training samples.

1.) Calculations performed with the training samples:

a) order the n training samples

$$y_1 < y_2 < \ldots < y_n \qquad \text{(IV.47)}$$

b) estimate k bin boundaries

$$\left(\hat{\xi}_{\frac{1}{k+1}} = y_{[\frac{n}{k+1}]+1}, \hat{\xi}_{\frac{2}{k+1}} = y_{[\frac{2n}{k+1}]+1}, \ldots\right.$$

$$\left., \hat{\xi}_{\frac{j}{k+1}} = y_{[\frac{jn}{k+1}]+1}, \ldots, \hat{\xi}_{\frac{k}{k+1}} = y_{[\frac{kn}{k+1}]+1}\right) . \qquad \text{(IV.48)}$$

By storing the k bin boundaries, the entire density estimate is stored so that f(x) can be estimated at a later time on line.

2.) Calculations performed to find $\hat{f}(x)$ from the bin boundaries in equation (IV.48):

a.) find $\hat{\xi}_{\frac{j}{k+1}}$ and $\hat{\xi}_{\frac{j+1}{k+1}}$ such that

$$\hat{\xi}_{\frac{j}{k+1}} \leq x \leq \hat{\xi}_{\frac{j+1}{k+1}}$$

b.) then

$$\hat{f}(x) = \begin{cases} 0 & x < a \\ \frac{1}{k+1} \Big/ (\hat{\xi}_{\frac{1}{k+1}} - a) & a \leq x \leq \hat{\xi}_{\frac{1}{k+1}} \\ \frac{1}{k+1} \Big/ (\hat{\xi}_{\frac{j+1}{k+1}} - \hat{\xi}_{\frac{j}{k+1}}) & \hat{\xi}_{\frac{j}{k+1}} \leq x \leq \hat{\xi}_{\frac{j+1}{k+1}} \\ \frac{1}{k+1} \Big/ (b - \hat{\xi}_{\frac{k}{k+1}}) & \hat{\xi}_{\frac{k}{k+1}} \leq x \leq b \\ 0 & x > b \end{cases} \qquad \text{(IV.49)}$$

IV.3 Comparison of Random Bin Density Estimate with Other Estimates

Section III.1 stated that density estimates generally are of the form $\frac{p(x \in \text{interval } \Delta)}{\text{width of } \Delta}$, and either the denominator is specified and the numerator estimated or the numerator is specified and the denominator estimated. The random bin model is of the latter type and so is similar to the NN estimate. Both estimate the interval width from the training samples. The random bin model is similar to the fixed bin model in that it is a step-function. In the random bin and fixed bin models, the density function is estimated for all x at once, and the total estimate is stored. Table IV.1 lists the properties of the random bin model and the three models discussed in Chapter III. Table IV.1 is similar to Table III.1 with the random bin estimate added. The remainder of this chapter discusses the estimates in more detail.

IV.3.1 Storage and Computation Requirements of Density Estimates

The storage and computation requirements of a density estimate can be divided into two parts. One part, to be called on-line, is for the storage of the data needed at the time of a classification decision and the amount of calculations required to make the decision. The other part, called off-line, is for any preprocessing that may be necessary before the data is stored for later use in a classification algorithm.

As an example of how off-line and on-line storage and processing might be utilized in practice, consider the EEG signals discussed in Section I.2. A possible decision problem is to determine from EEG

		Factor 2		Properties Influenced by Factor 1			
				In $f(x)=p(x\epsilon\Delta)/\Delta$			
		Total Point Estimate[1]	Single Point Estimate[2]	denominator specified, numerator estimated	numerator specified, denominator estimated	Difficulty of bin size choice[4] more / less	Convergence conditions as # training samples→∞
Factor 1	Bin Width Set by Training Samples	Random Bin	NN 17,18[3]		✓	✓	specified bin width→0 at such a rate that # samples in bin→∞
	Bin Width Specified	Fixed Bin 19,20 21,22,23	Parzen 16,18, 24	✓		✓	# samples specified in bin→∞ at such a rate that bin width→0[5]

Properties Influenced by Factor 2			
Is bin centered at x?	no	✓	
	yes		✓
Storage requirement	small	✓	
	large		✓
Computational complexity for any x	less	✓	
	more		✓
Accuracy in deterministic sense	less	✓	
	more		✓
Tail region problem	yes	✓	
	no		✓

1. In Total Point Estimate, the density function is estimated for all x at once, and the total estimate is stored.

2. In Single Point Estimate, all training samples are stored and the density is estimated separately for each x.

3. These numbers indicate references in the bibliography.

4. When the bin width is specified, there is a problem of how to choose it initially so as to contain a number of training samples that would give a reasonable estimate. In letting the training samples set the bin width, a reasonable estimate is more readily obtained.

5. The number of samples specified in the bin→∞ but a rate sufficiently slower than the total number of training samples→∞ in order that the bin width that contains the specified number of samples→0.

TABLE IV.1 Properties of Fixed Bin, Parzen, NN, and Random Bin Density Estimates

measurements the state of consciousness of a patient undergoing surgery. The information on the patient's state of consciousness would determine the amount of anesthetic to give the patient. Calculations on the EEG measurements to determine the density functions necessary for such a decision could be performed off-line before the surgery when large computer facilities would be available. During the operation, the testing on the patient's state of consciousness could be done on-line with small information storage facilities being required.

When the density estimate is a step-function calculated for all x at once as in the random bin and fixed bin models, off-line processing is necessary. But the on-line storage requirement of these estimates is small as only the bin boundaries and step-function values need be stored, and the on-line calculation of the density estimate for any x is simple because only the bin in which x lies needs be found. The Parzen and NN models have no off-line processing. But the on-line storage requirement is large since all training samples are stored, and more on-line calculations are required as the bin is centered at x every time an estimation is made.

As mentioned in the previous paragraph, the random and fixed bin estimates require off-line storage. In the fixed bin model, the fraction of training samples in each bin is calculated, and each training sample may be discarded onee the bin in which it lies has been found. In the random bin model, the training samples are ordered

and all samples must be stored during the placement of the bin boundaries. Thus the off-line processing requirement of the random bin model is larger than that of the fixed bin model. The fixed bin estimate is also easier to update with additional samples.

IV.3.2 Bin Placement

In the random bin and NN models, the interval positions are determined by the training samples, while in the fixed bin and Parzen models, the interval positions are specified before training samples are known. When the intervals are specified beforehand, a bin may contain a very high proportion of the samples; it may be necessary to change the interval and estimate the density again to increase the accuracy.

The centering of the bin at x in the NN and Parzen estimates provides more deterministic accuracy. The random and fixed bin models do not center their bins at x, but the decreased deterministic accuracy is traded for smaller on-line storage and processing requirements.

Properties of the random and fixed bin models can be combined into one estimate. The bin boundaries could be placed by some of the training samples, then the bins could be taken as specified and the fixed bin method applied to the other samples. Such a mixed estimate would combine the two modes of density estimation, which are either specifying the denominator of $\frac{p(x \in \text{interval } \Delta)}{\text{width of } \Delta}$ and estimating the numerator or specifying the numerator and estimating the denominator. The mixed density estimate would operate in each mode one at a time.

Sebestyen and Edie [27] have formulated a density estimate that is one possible way of combining the two modes of density estimation mentioned in the previous paragraph. Sebestyen and Edie determine both the number of bins and bin sizes from the training samples. The estimate is a step-function. First, an initial set of bins is chosen. Then by applying the training samples, some bins are enlarged and some reduced, and some new bins are created and some old ones combined. The flat parts of the density function are approximated by a few, large bins, and the rapidly varying parts by more, smaller bins. The motivation of the estimate is to minimize the mean square error

$$\int_{-\infty}^{\infty} (f(x)-\hat{f}(x))^2 dx \qquad \text{(IV.50)}$$

and require little storage.

Figures III.4 a,b, and c show an illustrative comparison of the estimates of Sebestyen and Edie, fixed bin, and random bin. The Sebestyen and Edie method appears to come the closest to minimizing the mean square error. But since the density function estimate is to be used to classify observations, it seems more appropriate that the estimate should have greater accuracy where observations are more likely to occur. In other words, the estimate should be more accurate nearer the peaks of the density. Rather than trying to minimize the mean square error, a more appropriate criterion is to minimize

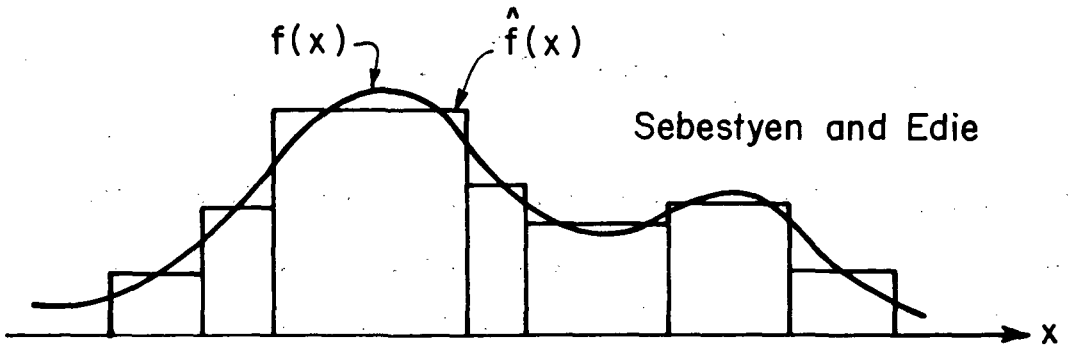


FIG. IV. 4a

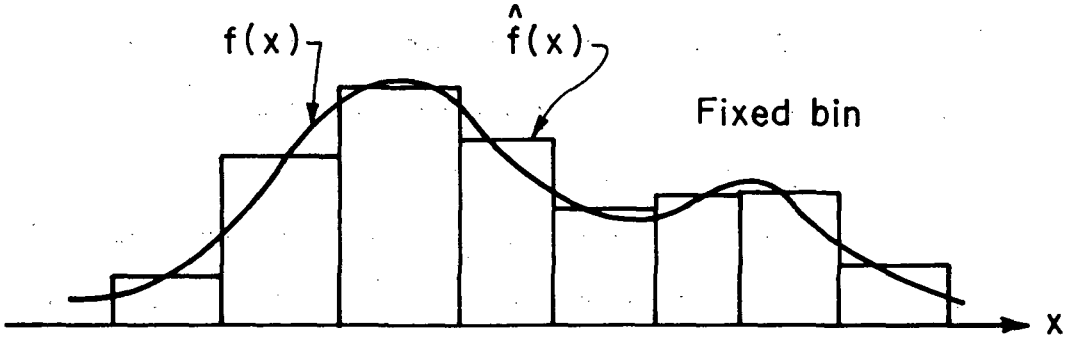


FIG. IV. 4b

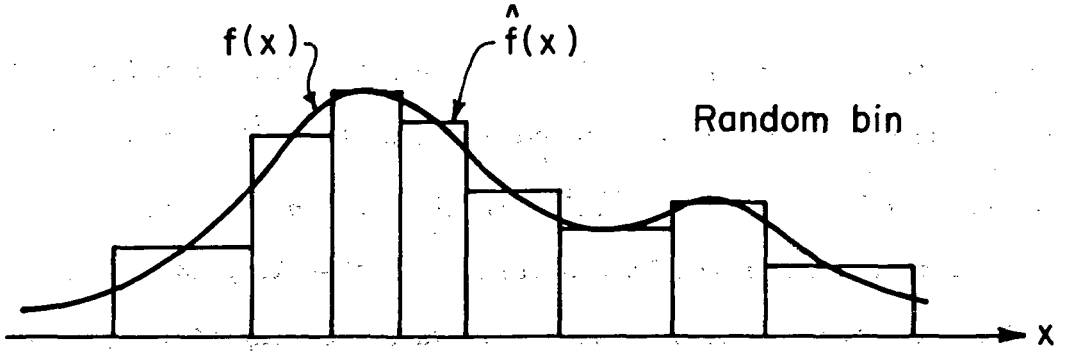


FIG. IV. 4c

Comparison of Density Estimates of Sebestyen and Edie, Fixed Bin, and Random Bin

$$\int_{-\infty}^{\infty} (f(x)-\hat{f}(x))^2 f(x)dx \quad . \qquad \text{(IV.51)}$$

Equation (IV.51) weighs more heavily the higher values of the density function where more observations are likely to occur. Since the random bin model places the bins so each bin contains approximately the same number of training samples, more bins are concentrated where more samples occur and the model comes closer to satisfying equation (IV.51). It is of course possible to vary the random bin model as presented in this thesis and to specify different numbers of training samples for different bins.

IV.3.3 Tail Region Problem

A problem arises with step-function estimates when the random variable X is distributed over the interval $(-\infty,\infty)$. If $f(x)$ is estimated for x less than the lowest bin boundary or greater than the highest bin boundary, $f(x)$ will be zero. For example in the random bin model in equation (IV.21), $\hat{f}(x) = 0$ for $x < \hat{\xi}_{\frac{1}{k+1}}$ or $x > \hat{\xi}_{\frac{k}{k+1}}$. Figure IV.5 illustrates this occurance. If an estimate of $f(x)$ is all that is desired in the tail regions, then $\hat{f}(x) = 0$ is a reasonable estimate. A problem occurs when $\hat{f}(x)$ becomes part of an estimated likelihood ratio $\hat{f}(x|C^2)/\hat{f}(x|C^1)$ as is the case in the estimated version of the Wald sequential probability ratio test to be presented in the next chapter. When a string of t observations has been taken and x_t results in either $\hat{f}(x_t|C^1) = 0$ or $\hat{f}(x_t|C^2) = 0$, the likelihood ratio of the t observations will be zero or infinity, and

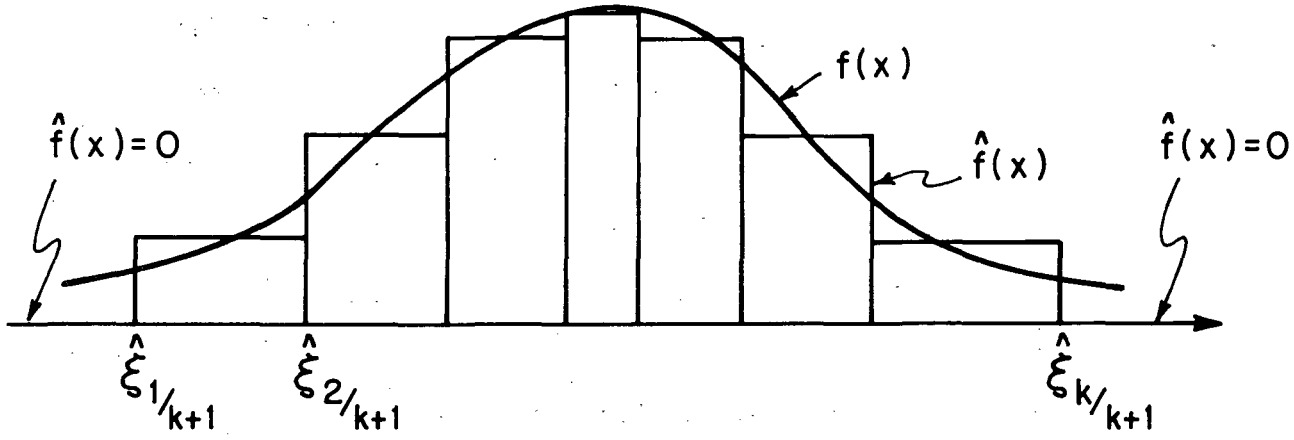


FIG. IV.5

Tail Regions of Random Bin Density Estimate

will cause a decision to be made immediately regardless of the previous observations. This phenomenon leads to more error decisions in the sequential test to be presented than should be allowed by the specific error probabilities. The reason is that a decision is made on the basis of only the one observation. The likelihood ratio ignores previous observations, and the test does not evaluate enough observations for the error rates to be small. This tail region problem, as it will be called in this thesis, is discussed further in Chapter V when the estimated version of the likelihood ratio is presented.

The Parzen and NN models avoid the tail region problem since their density estimates are continuous in x.

IV.3.4 Conclusion to Comparison of Density Estimates

Table IV.1 is again recommended for a comparison of the various estimates. The next chapter explores the use of the random bin estimate in an estimated SPRT. The random bin model is chosen because of its small on-line storage and processing requirements and its placing of the interval widths by the training samples.

Appendix IV.1 - Discussion of Convergence Proofs of Density Estimates

This appendix discusses some factors involved in showing convergence of density estimates. Parzen [16] shows convergence of his estimate in the mean square sense. Loftsgaarden and Quesenberry [17] show convergence in probability, and this report shows convergence of the random bin density estimate in probability. Mean square convergence is a stronger form of convergence, and in fact it implies convergence in probability. The reason that convergence of the NN and random bin models has been shown in probability appears to be that their structure makes convergence harder to prove (it should be noted that it has not been shown that they do not converge in the mean square sense or with probability one).

The basic form of a density approximation is

$$\frac{p(\text{observation } \varepsilon\ \Delta)}{\Delta} \quad . \qquad \text{(IV.1.1)}$$

As mentioned in Section III.2, a density estimate can either specify the denominator and estimate the numerator or specify the numerator and estimate the denominator. The Parzen model estimates the numerator, and the NN and random bin models estimate the denominator. Because of this, it is more difficult to find the means and variance of the NN and random bin models. Estimating the denominator of equation (IV.1.1) means estimating the interval width that contains a specified fraction of the training samples. Distributions of order statistics are involved, and it is difficult to calculate the variance of interval

width estimates from the densities of order statistics. Estimating the numerator of equation (IV.1.1) involves estimating F(x+h) - F(x-h), which has a variance that is easier to find.

To illustrate the factors discussed in the preceding paragraph, some examples will be given of the type of calculations involved for finding the variances of density estimates. Let the density estimates be based on $X_1, X_2, \ldots, X_n$ where $X_1, X_2, \ldots, X_n$ are independent random variables identically distributed as the random variable X with absolutely continuous distribution function F(x) and with probability density function f(x).

The first density estimate Parzen considers is

$$\hat{f}_{Parzen}(x) = \frac{S_n(x+h)-S_n(x-h)}{2h} \qquad \text{(IV.1.2)}$$

where $S_n(x)$ is the fraction of samples less than x. The covariance of $S_n(x)$ and $S_n(x')$ is [14]

$$\text{cov}(S_n(x),S(x')) = \frac{1}{n}\,[F(\min(x,x'))- F(x)F(x')] \ .$$

For the general Parzen estimate

$$\hat{f}_{Parzen}(x) = \frac{1}{nh} \sum_{j=1}^{n} K\left(\frac{x-x_i}{h}\right) , \qquad \text{(IV.1.3)}$$

Parzen shows that

$$\lim_{n\to\infty} nh\ \text{var}[f_n(x)] = f(x) \int_{-\infty}^{\infty} K^2(y)dy \ . \qquad \text{(IV.1.4)}$$

It is evident that the limit of the variance of Parzen estimate can

be found and mean square convergence can be shown.

The NN density estimate is

$$\hat{f}_{NN}(x) = \frac{\ell-1}{n} \, /2|x-x_\ell| \qquad \text{(IV.1.5)}$$

where x_ℓ is the ℓ-th nearest sample to x. The NN estimate involves order statistic theory. In making calculations on the NN estimate, the type of density function involved is that of the k-th largest sample y_k whose density is

$$\frac{n!}{(k-1)!(n-k)!} \, [1-F(y_k)]^{n-k}[F(y_k)]^{k-1}f(y_k) \quad . \qquad \text{(IV.1.6)}$$

The random bin estimate is

$$f_{\substack{\text{Random} \\ \text{bin}}}(x) = \frac{1}{k+1} \, / \, (\hat{\xi}_{\frac{j+1}{k+1}} - \hat{\xi}_{\frac{j}{k+1}}) \quad \text{for} \quad \hat{\xi}_{\frac{j}{k+1}} < x < \hat{\xi}_{\frac{j+1}{k+1}} \qquad \text{(IV.1.7)}$$

where $\hat{\xi}_p = y_{[np]+1}$ is the estimate of the p-th order quantile ξ_p. The random bin estimate also involves order statistic theory, and the type of density function used for making calculations on the random bin estimate is that of the joint density of $\hat{\xi}_p$ and $\hat{\xi}_q$, which is

$$\frac{n!}{(i-1)!(j-i-1)!(n-j)!} \, [F(\hat{\xi}_p)]^{i-1}[F(\hat{\xi}_q)-F(\hat{\xi}_p)]^{j-i-1}$$

$$\cdot[1-F(\hat{\xi}_q)]^{n-j}f(\hat{\xi}_p)f(\hat{\xi}_q) \qquad \text{(IV.1.8)}$$

where $p < q$, $i = [np]+1$, and $j = [nq]+1$.

Since it is difficult to find explicit expressions for the variance of random variables with density functions in equations (IV.1.6) and (IV.1.8) and with $F(x)$ and $f(x)$ unknown, explicit expressions for the variance of the NN and random bin estimate are even more difficult to find. Thus the convergence of the NN and random bin estimates has been shown in probability by methods that do not involve finding explicit expressions for the variance of the estimates, such as using asymptotic distributions.

CHAPTER V

ESTIMATED SPRT

Chapter IV developed a density function estimate with the intent of utilizing it in a classification procedure. This chapter discusses the Wald sequential probability ratio test (SPRT) and then forms an estimated SPRT with the random bin density estimate. The SPRT has been chosen since the decision problem involving the EEG responses discussed in Section I.2 is particularly well suited for a sequential test. Also a SPRT with density estimates presents some additional interesting problems which occur only infrequently in tests that decide on the basis of only one observation such as the Bayes decision rule. Some of these problems that will be investigated in this report are estimating densities in the tail regions and estimating densities of dependent observations.

V.1 Review of SPRT

A well-known sequential test is the Wald SPRT [4,28,29]. In the SPRT, the error probabilities are specified

$$\alpha \triangleq p(\text{error of type I}) = p(\text{decide } C^2 | C^1 \text{ true})$$
$$\beta \triangleq p(\text{error of type II}) = p(\text{decide } C^1 | C^2 \text{ true}) \qquad \text{(V.1)}$$

Define the likelihood ratio of t observations

$$L(x_1, x_2, \ldots, x_t) = \frac{f(x_1, x_2, \ldots, x_t | C^2)}{f(x_1, x_2, \ldots, x_t | C^1)} \qquad \text{(V.2)}$$

and two thresholds

$$A = \frac{1-\beta}{\alpha} , \qquad B = \frac{\beta}{1-\alpha} . \qquad \text{(V.3)}$$

The operation of the SPRT is as follows.

1.) Take the first observation x_1. If

$$L(x_1) \leq B \qquad \text{decide } C^1$$

$$B < L(x_1) < A \qquad \text{observe the next observation } x_2$$

$$L(x_1) \geq A \qquad \text{decide } C^2$$

2.) If another observation is taken, say the t-th observation x_t,

$$L(x_1,x_2,\ldots,x_t) \leq B \qquad \text{decide } C^1$$

$$B < L(x_1,x_2,\ldots,x_t) < A \qquad \text{observe the next observation } x_{t+1}$$

$$L(x_1,x_2,\ldots,x_t) \geq B \qquad \text{decide } C^2$$

3.) Repeat step 2 on the next observation until a decision is made.

The SPRT takes new observations until the information contained in the string of observations is sufficient that the probabilities of type I and type II errors in making a decision are equal to the specified values α and β respectively. The SPRT has the property that among all tests for which α and β are specified, the SPRT requires the smallest number of observations, on the average, to reach a decision [2,29].

When the observations x_i are independent, the likelihood ratio can be written as

$$L(x_1,x_2,\ldots,x_t) = \frac{f(x_1|C^2)f(x_2|C^2)\cdots f(x_t|C^2)}{f(x_1|C^1)f(x_2|C^1)\cdots f(x_t|C^1)} \qquad \text{(V.5)}$$

For convenience, in the remainder of the report $f(x|C^1)$ is written $f_1(x)$ and $f(x|C^2)$ is written $f_2(x)$.

The SPRT obtains the information contained in a string of observations by evaluating the density functions of each class at the observation

values. Knowledge of the density functions of each class are required for the SPRT, and so the test is not directly applicable to the case where the only prior knowledge is that of training sets.

Fu [30] has developed a partially distribution-free version of the SPRT that uses the training samples of only one class, say C^1. If the samples from C^1 have an arbitrary distribution function $F(x)$, then the samples from C^2 are assumed to have the Lehman alternative distribution, which is $F^r(x)$, $r>0$. After each observation from the unknown class is taken, two sets of samples are formed -- one from samples of C^1 and the other by alternating samples of C^1 with observations from the unknown class. The samples of both sets are ordered, and the density functions of the two ordered sets are found. By assuming the distributions of C^1 and C^2 are $F(x)$ and $F^r(x)$ respectively, the ratio of the densities of the two orderings is independent of $F(x)$. This new ratio of densities is used in the SPRT to determine if the second ordering contains only samples from C^1 or samples from both C^1 and C^2. Fu has used training samples from only one class and has assumed the distribution of C^2 is $F^r(x)$, $r > 0$, where $F(x)$ is the distribution of C^1.

The method presented in this chapter uses training samples from both classes and forms an estimated likelihood ratio for use in the SPRT from estimates of the density functions of each class. The method is distribution-free in that it does not require any knowledge of $f_1(x)$ and $f_2(x)$.

V.2 Random Bin Estimate in SPRT

V.2.1 Presentation of Random Bin Estimate in SPRT

Since the density functions $f_1(x)$ and $f_2(x)$ are unknown, they can be estimated from training samples of each class, and an estimated

likelihood ratio can be formed

$$\hat{L}(x_1,x_2,\ldots,x_t) = \frac{\hat{f}_2(x_1)\hat{f}_2(x_2)\cdots\hat{f}_2(x_t)}{\hat{f}_1(x_1)\hat{f}_1(x_2)\cdots\hat{f}_1(x_t)} \quad . \tag{V.6}$$

Let

n_1 be the number of training samples in class 1,

n_2 be the number of training samples in class 2,

k_1 be the number of quantiles for $\hat{f}_1(x)$,

k_2 be the number of quantiles for $\hat{f}_2(x)$,

$\hat{\xi}_{\frac{j}{k_1+1}}$, $j=1,2,\ldots,k_1$ be the k_1 quantiles for $\hat{f}_1(x)$, and

$\hat{\eta}_{\frac{j}{k_2+1}}$, $j=1,2,\ldots,k$ be the k_2 quantiles for $\hat{f}_2(x)$. (V.7)

The estimate of $L(x_1,x_2,\ldots,x_t)$ formed from the random bin density estimate is

$$\hat{L}(x_1,x_2,\ldots,x_t) = \frac{\hat{f}_2(x_1)\hat{f}_2(x_2)\cdots\hat{f}_2(x_t)}{\hat{f}_1(x_1)\hat{f}_1(x_2)\cdots\hat{f}_1(x_t)} \tag{V.8}$$

where

$$\hat{f}_1(x_i) = \frac{1}{k_1+1} \Big/ \left(\hat{\xi}_{\frac{j+1}{k_1+1}} - \hat{\xi}_{\frac{j}{k_1+1}}\right) \quad \text{for} \quad \hat{\xi}_{\frac{j}{k_1+1}} \leq x_i \leq \hat{\xi}_{\frac{j+1}{k_1+1}} \tag{V.9}$$

and

$$\hat{f}_2(x_i) = \frac{1}{k_2+1} \Big/ \left(\hat{\eta}_{\frac{\ell+1}{k_1+1}} - \hat{\eta}_{\frac{\ell}{k_1+1}}\right) \quad \text{for} \quad \hat{\eta}_{\frac{\ell}{k_1+1}} < x_i < \hat{\eta}_{\frac{\ell+1}{k_2+1}} \tag{V.10}$$

If $k_1 = k_2 = k$, the computation of $\hat{L}(x_1,x_2,\ldots,x_t)$ is reduced since

$$\frac{\hat{f}_2(x_i)}{\hat{f}_1(x_i)} = \left(\hat{\xi}_{\frac{j+1}{k_1+1}} - \hat{\xi}_{\frac{j}{k_1+1}}\right) \Big/ \left(\hat{\eta}_{\frac{\ell+1}{k_2+1}} - \hat{\eta}_{\frac{\ell}{k_2+1}}\right)$$

where $\hat{\xi}_{\frac{j}{k_1+1}} \leq x_i \leq \hat{\xi}_{\frac{j+1}{k_1+1}}$ and $\hat{\eta}_{\frac{\ell}{k_2+1}} \leq \leq \hat{\eta}_{\frac{\ell+1}{k_2+1}}$. (V.11)

Since $\hat{f}_i(x_1)$, $\hat{f}_i(x_2),\ldots,\hat{f}_i(x_t)$ are estimated from the same training samples, $\hat{f}_i(x_1)$, $\hat{f}_i(x_2),\ldots,\hat{f}_i(x_t)$ are in general dependent, i=1,2. So

$$E[\hat{f}_i(x_1)\hat{f}_i(x_2)\cdots\hat{f}_i(x_t)] \neq E\hat{f}_i(x_1)E\hat{f}_i(x_2)\cdots E\hat{f}_i(x_t), \quad \text{for } i=1,2 \qquad (V.12)$$

and $\hat{L}(x_1,x_2,\ldots,x_t)$ is a biased estimator of $L(x_1,x_2,\ldots,x_t)$. But the next section shows that $\hat{L}(x_1,x_2,\ldots,x_t)$ converges in probability to $L(x_1,x_2,\ldots,x_t)$ as $n_1 \to \infty$ and $n_2 \to \infty$ and is thus a consistent estimate (see also conclusion to this chapter).

V.2.2 Convergence of Likelihood Ratio

Theorem 3: Let $X_1^1,X_2^1,\ldots,X_{n_1}^1$, be a set of independent random variables identically distributed as the random variable X^1 with absolutely continuous distribution function $F_1(x)$ and with probability density function $f_1(x)$, and let $X_1^2,X_2^2,\ldots,X_{n_2}^2$ be a set of independent random variables distributed as X^2 with $F_2(x)$ and $f_2(x)$ similarly defined. Let $\hat{f}_1(x)$ be an estimate $f_1(x)$ and $\hat{f}_2(x)$ an estimate of $f_2(x)$

where the estimates are defined in Theorem 2 of Chapter IV. Define

$$\hat{L}(y_1,y_2,\ldots,y_t) = \frac{\hat{f}_2(y_1)\hat{f}_2(y_2)\cdots\hat{f}_2(y_t)}{\hat{f}_1(y_1)\hat{f}_1(y_2)\cdots\hat{f}_1(y_t)}$$

Then $\hat{L}(y_1,y_2,\ldots,y_t)$ converges in probability to

$$L(y_1,y_2,\ldots,y_t) = \frac{f_2(y_1)f_2(y_2)\cdots f(y_t)}{f_1(y_1)f_1(y_2)\cdots f(y_t)}$$

as $n_1 \to \infty$ and $n_2 \to \infty$ for all $y_1,y_2,\ldots,y_t$ in the neighborhood of which $f_1(x), f_1'(x), f_2(x)$ and $f_2'(x)$ are continuous and $f_1(x) \neq 0$ and $f_2(x) \neq 0$.

Proof: From Theorem 2, $\hat{f}_1(y_i)$ converges in probability to $f_1(y_i)$ as $n_1 \to \infty$ and $\hat{f}_2(y_i)$ converges in probability to $f_2(y_i)$ as $n_2 \to \infty$. The proof of the corollary follows directly from the theorem from Krickeberg [31] that if the sequences of random variables $\xi_n, \eta_n, \ldots, \rho_n$ converge in probability to $\xi, \eta, \ldots, \rho$ then the sequence $g(\xi_n,\eta_n,\ldots\rho_n)$ converges in probability to $g(\xi,\eta,\ldots,\rho)$ if g is a continuous function and $g(\xi,\eta,\ldots,\rho)$ is finite.

This section has proposed an estimated SPRT where the likelihood ratio is formed from random bin density estimates of each class. The estimated likelihood ratio of independent observations was shown to converge in probability to the true likelihood ratio. The remainder of this chapter discusses the application of the SPRT to classification problems.

V.3 Tail Region Estimation Problem in the Random Bin SPRT

One difficulty that occurs with a step-function density estimate such as the random bin model is the estimation of the tail regions of the density function. As an example of this problem, consider two overlapping density functions as illustrated in Figure V.1 with their possible estimates in Figure V.2. Assume that an estimated SPRT is being performed and that after t observations no decision has been made. Thus

$$B < \hat{L}(x_1,x_2,\ldots,x_t) < A \quad .$$

Suppose further that the observations to be classified belong to class 1 and that the (t+1)-th observation is greater than $\hat{\xi}_{\frac{k}{k+1}}$.

This means that $\hat{f}_1(x_{t+1}) = 0$ and so

$$\hat{L}(x_1,x_2,\ldots,x_t,x_{t+1}) = \hat{L}(x_1,x_2,\ldots,x_t)\,\frac{\hat{f}_2(x_{t+1})}{\hat{f}_1(x_{t+1})} = \infty < A \ .$$

A wrong decision that the observations belong to class 2 is made. Since $\hat{f}_1(x_{t+1}) = 0$ for any $x_{t+1} > \hat{\xi}_{\frac{k}{k+1}}$, a decision of class 2 is made for any $x_{t+1} > \hat{\xi}_{\frac{k}{k+1}}$. However, if the actual density functions are known, it is possible that the ratio $\frac{f_2(x_{t+1})}{f_1(x_{t+1})} > B$ for $x_{t+1} > \hat{\xi}_{\frac{k}{k+1}}$ and that the ratio $f_2(x_{t+1})/f_1(x_{t+1})$ is sufficiently small so

$$B < L(x_1,x_2,\ldots,x_t,x_{t+1}) = L(x_1,x_2,\ldots,x_t)\,\frac{f_2(x_{t+1})}{f_1(x_{t+1})} < A \ .$$

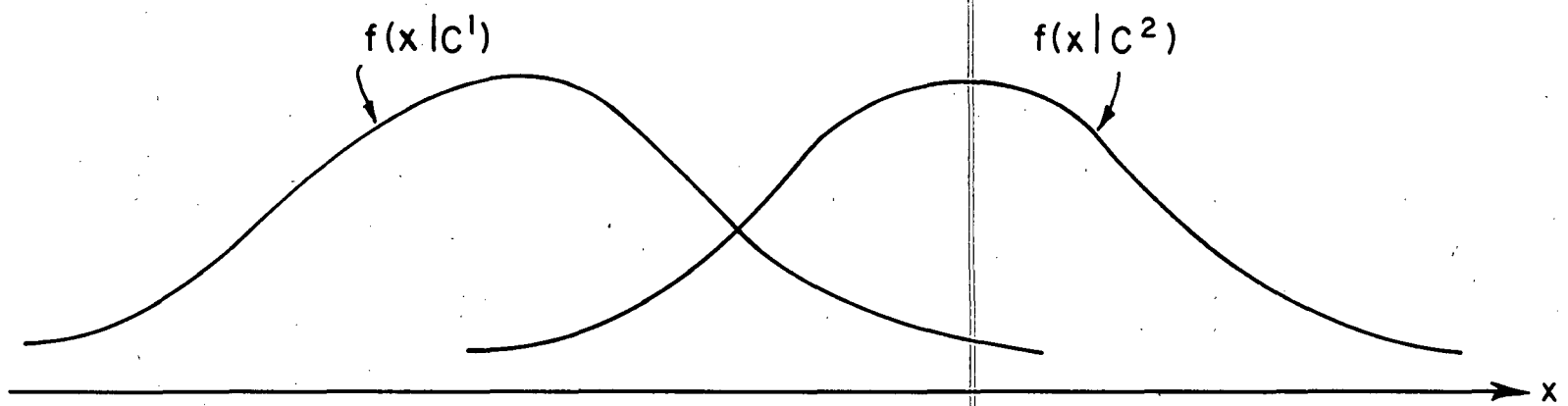


FIGURE V.1

Example of Two Overlapping Density Functions

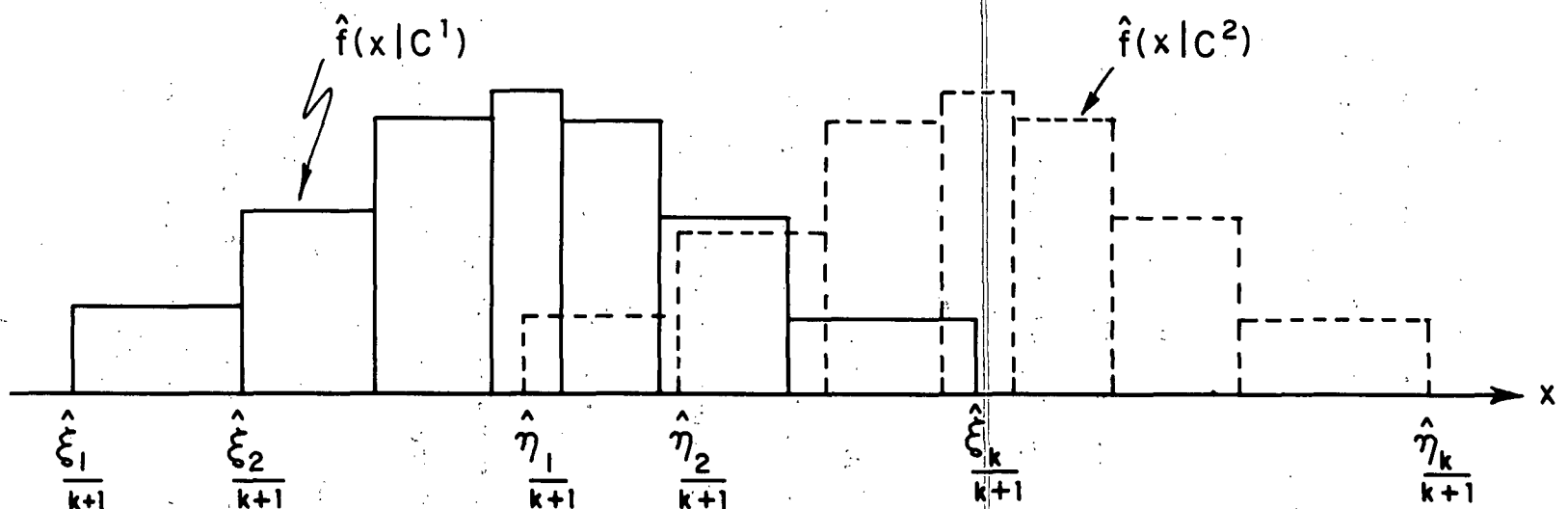


FIGURE V.2

Example of Two Overlapping Random Bin Density Estimates

Thus it is possible that after the (t+1)-th observation where $x_{t+1} > \hat{\xi}_{\frac{k}{k+1}}$ the estimated SPRT decides class 2 and the true SPRT makes no decision. Estimating the tail regions of a density function to be zero causes more classification errors than desired. When $\hat{f}_1(x_{t+1}) = 0$, a decision is based on only the one observation x_{t+1}; the information contained in the previous t observations is neglected. The same difficulty is encountered when classifying observations from class 2 that are less than $\hat{\eta}_{\frac{1}{k+1}}$. Experimental results appearing later in this chapter verify that the tail region problem does result in more classification errors than would be expected from the specified error probabilities. A step-function estimate does not cause excessive classification errors on observations between the tail regions since the likelihood ratio is not zero or infinity. Consequently a decision is not automatically made from the information supplied by the one observation.

The tail region problem occurs mainly when several observations are considered at once. If a classification process decides on the basis of only one observation, such as the Bayes decision rule, then estimating the tail regions to be zero may be acceptable. Since no information additional to the one observation is to be taken, no information is ignored by the likelihood ratio being zero or infinity.

Two techniques for handling the tail region problem are discussed in the next few sections. The methods either estimate the tail regions differently or vary the SPRT.

V.3.1 Requiring Several Observations to Fall in the Tail Regions

One solution to the tail region problem that has worked experimentally treats the observations from the tail regions separately from the likelihood ratio. The method makes a decision of class 2 if r observations fall greater than $\hat{\xi}_{\frac{k}{k+1}}$, refer to Figure V.2, and a decision of class 1 if r observations fall less than $\hat{\eta}_{\frac{1}{k+1}}$. Only observations between $\hat{\eta}_{\frac{1}{k+1}}$ and $\hat{\xi}_{\frac{k}{k+1}}$ are included in the likelihhod ratio. A decision about a string of observations is made in one of two ways, either by the likelihood ratio of observations between $\hat{\eta}_{\frac{1}{k+1}}$ and $\hat{\xi}_{\frac{k}{k+1}}$ falling outside the thresholds A and B, or by the number of observations less than $\hat{\eta}_{\frac{1}{k+1}}$ equaling r or the number of observations greater than $\hat{\xi}_{\frac{k}{k+1}}$ equaling r.

The motivation for this solution to the tail region problem is that more observations are used in the decision process if r observations rather than one are required to fall in each tail region before deciding. With an increase in the required number of observations, the decision is more likely to be made by the SPRT rather than the tail region test, and the combined test is likely to be more accurate. The error rate is decreased by increasing r, but the average number of observations required for a decision is increased. If r is made very large, the observations in the tail regions do not contribute at all to the decision process.

A disadvantage of the technique presented in this section is that the tail region treatment departs from the likelihood ratio method of the SPRT. Since observations below $\hat{\eta}_{\frac{1}{k+1}}$ or above $\hat{\xi}_{\frac{k}{k+1}}$ are not included in the SPRT structure, the error probabilities of the altered test may differ from those specified in a standard SPRT. The next section presents a method that estimates the tail regions with a different density estimate and preserves the SPRT structure for all observations.

V.3.2 NN Tail Region Estimate

Another way of handling the tail region problem is to employ the nearest neighbor (NN) density estimate of Loftsgaarden and Quesenberry explained in Section III.4.1. The NN estimate is

$$\hat{f}(x) = \frac{\ell(n)-1}{n} \Big/ 2|x-x_{\ell(n)}| \qquad (V.13)$$

where n is the number of training samples and $x_{\ell(n)}$ is the $\ell(n)$-th nearest training sample to x according to the distance measure $|x-y|$. This estimate is continuous in x and tends to zero only as x approaches infinity. Disadvantages of the estimate are that all training samples must be stored and the $\ell(n)$-th nearest sample to x must be found for each x.

The NN estimate, however, can be used to advantage in the tail regions. For any observation x below a certain value, the same training sample is always the ℓ-th nearest sample to x, and the same is true for any x exceeding a certain value. Figure V.3 provides an illustration

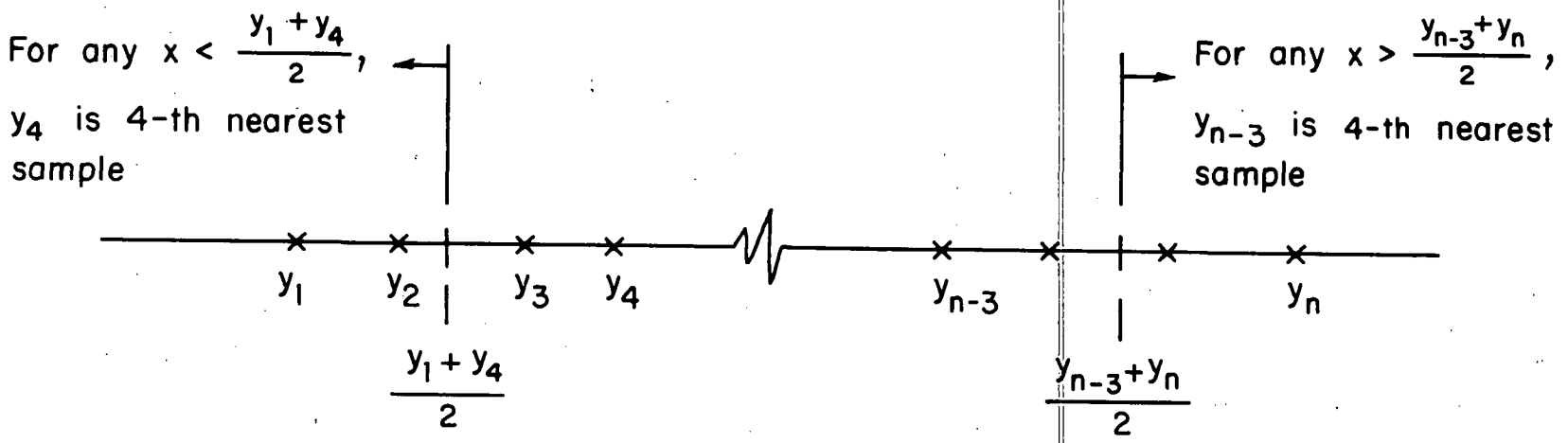


FIGURE V.3

Nearest Neighbor Density Estimate for Tail Regions

with $\ell = 4$. Let $y_1, y_2, \ldots, y_n$ be a set of n ordered training samples. For any x less than the midpoint between y_1 and y_ℓ, y_ℓ is always the ℓ-th nearest training sample to x. So the NN density estimate for any observation $x < \frac{y_1+y_\ell}{2}$ is

$$\hat{f}(x) = \frac{\ell-1}{n} \Big/ 2|x-y_\ell| \tag{V.14}$$

The estimate in equation (V.14) is greater than zero in the tail regions. The values of y_ℓ and the midpoint of y_1 and y_ℓ are the only information that needs to be stored for later use of the density estimates of the tail regions. At the upper tail of the density, $y_{n+1-\ell}$ is always the ℓ-th nearest sample to any $x > (y_{n+1-\ell}+y_n)/2$. So

$$\hat{f}(x) = \frac{\ell-1}{n} \Big/ 2|x-y_{n+1-\ell}| \quad \text{for } x > \frac{y_{n+1-\ell}+y_n}{2} \; . \tag{V.15}$$

The random bin density estimate with the NN tail region estimate is illustrated in Figure V.4. $\hat{\xi}_{\frac{1}{k+1}} = y_\ell$ where $\ell = [\frac{n}{k+1}] + 1$ (see equation (IV.20)) is the smallest bin boundary. For any $x < \tilde{a} \triangleq (y_1 + \hat{\xi}_{\frac{1}{k+1}})/2$, $\hat{\xi}_{\frac{1}{k+1}} = y_\ell$ is the ℓ-th nearest training sample to x. Similarly $\hat{\xi}_{\frac{k}{k+1}} = y_{n+1-\ell}$ where $\ell = n = [\frac{kn}{k+1}]$ is the largest bin boundary, and for $x > \tilde{b} \triangleq (\hat{\xi}_{\frac{k}{k+1}} + y_n)/2$, $\hat{\xi}_{\frac{k}{k+1}} = y_{n+1-\ell}$ is the ℓ-th nearest sample to x. The bins have been chosen so that each bin contains approximately $[\frac{n}{k+1}]$ training samples. Referring to Figure V.4 again, the density estimate still has not been determined for the regions

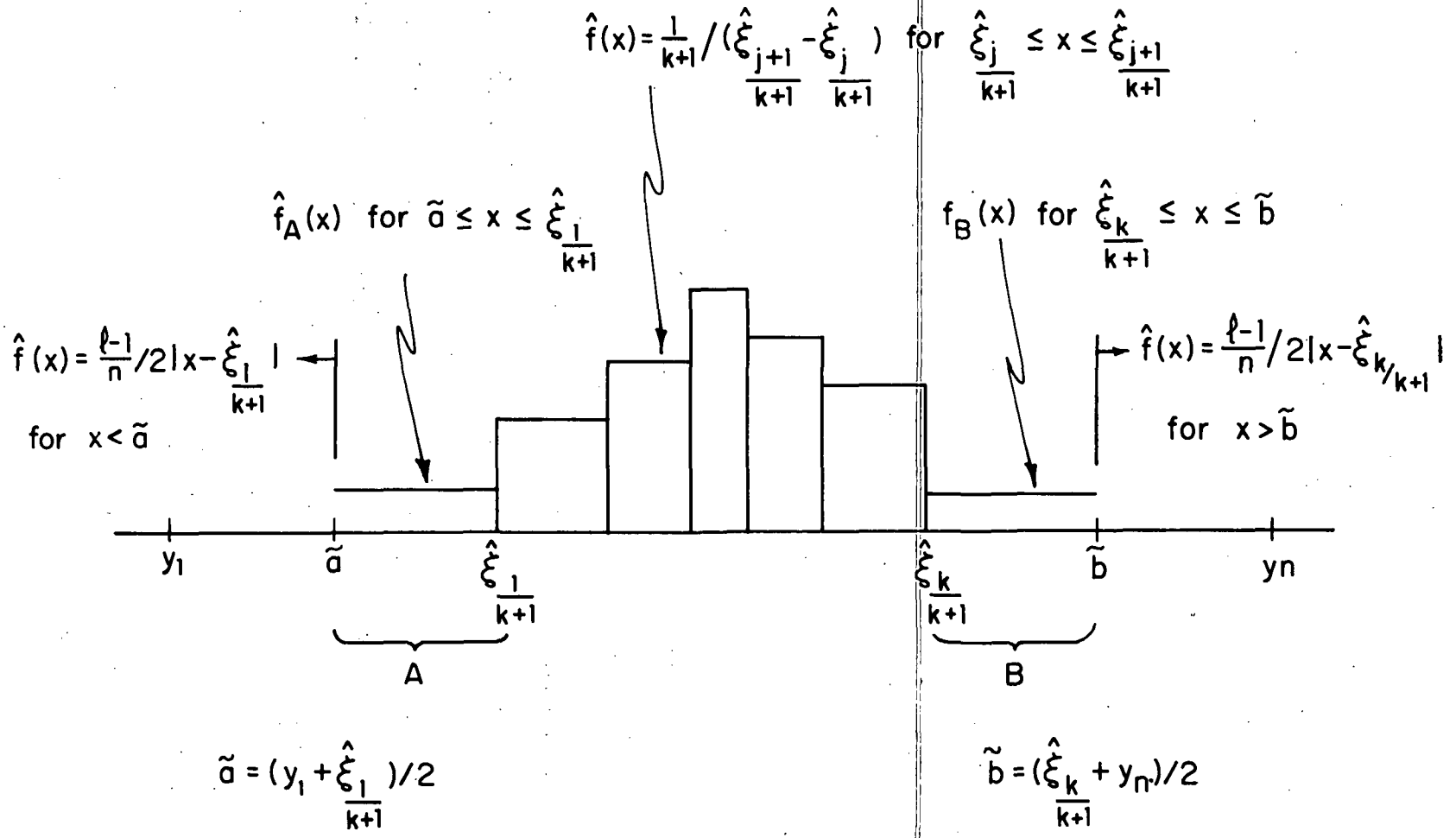


FIGURE V.4

Random Bin Density Estimate with NN Tail Region Estimate

$A = \{x | \tilde{a} < x < \hat{\xi}_{\frac{1}{k+1}}\}$ and $B = \{x | \hat{\xi}_{\frac{k}{k+1}} < x < \tilde{b}\}$. The density in regions A and B is estimated for the experimental examples in this report by centering a bin from the NN model at the midpoint of the regions. $(\tilde{a} + \hat{\xi}_{\frac{1}{k+1}})/2$ is the midpoint of region A, and $(\hat{\xi}_{\frac{k}{k+1}} + \tilde{b})/2$ of region B. Thus with each bin containing ℓ samples,

$$\hat{f}_A(x) = \frac{\ell-1}{n} / 2 \left| \frac{(\tilde{a} + \hat{\xi}_{\frac{1}{k+1}})}{2} - {}^{A}_{\ell}y \right| \quad \text{for } \tilde{a} < x < \hat{\xi}_{\frac{1}{k+1}} \tag{V.16}$$

where ${}^{A}_{\ell}y$ is the ℓ-th nearest training sample to $(\tilde{a} + \hat{\xi}_{\frac{1}{k+1}})/2$, and

$$f_B(x) = \frac{\ell-1}{n} / 2 \left| \frac{\hat{\xi}_{\frac{k}{k+1}} + \tilde{b}}{2} - {}^{B}_{\ell}y \right| \quad \text{for } \hat{\xi}_{\frac{k}{k+1}} < x < \tilde{b} \tag{V.17}$$

where ${}^{B}_{\ell}y$ is the ℓ-th nearest training sample to $(\hat{\xi}_{\frac{k}{k+1}} + \tilde{b})/2$.

The density has a constant value throughout each interval A and B. The tail regions were estimated by the NN model in the manner explained in order to assure that the bins in the tail regions contain approximately the same number of samples as the bins in the center region which had been estimated by the random bin model.

V.4 Experimental Results of the Estimated SPRT Tested on Gaussian Data

This section shows the results of the SPRT with the random bin density estimate tested on independent, scalar Gaussian samples. The mean of the distribution from class 1 is -0.8, and the mean of class 2

is +0.8. The variance of both classes is one. Experimentally, a good relationship between the number of training samples n and the number of quantiles k seems to be $k \approx n^{1/2}$. Loftsgaarden and Quesenberry [17] also state that on the basis of some emperical work using their estimate a value of ℓ near $n^{1/2}$ appears to give good results. For the following examples, n = 999 training samples and k = 29 quantiles (giving k+1 = 30 bins) were used for each density estimate. After the density functions of both classes are estimated, the estimates were tested in the SPRT with one thousand test observations from each class. The test was conducted with several values of the error probabilities, α= p(decide class 2|class 1 true) and β = p(decide class 1|class 2 true). The next two sections present the experimental results for the two tail region treatments discussed in Sections V.3.1 and V.3.2.

V.4.1 Experimental Results of the Estimated SPRT With r Observations Falling in the Tail Regions

Section V.3.1 discusses the treatment of the tail region where a decision is made either by the SPRT applied to observations between the tail regions or after r observations fall in one of the tail regions. Table V.1 shows the experimental results. Values of r from one to five were considered. The error rates in Table V.1 for r = 1 represent neglecting the tail region problem and allowing $\hat{f}_1(x)$ and $\hat{f}_2(x)$ to be zero for observations in the tail regions. It is observed that the experimental error rates for r = 1 are indeed higher than the specified α and β. The error rates are decreased by increasing r. More obser-

$\alpha = \beta$	Number observations in tail regions for decision r	Experimental Results			
		Experimental error rate		Experimental average number observations for decision	
		Class 1	Class 2	Class 1	Class 2
.1	1	.084	.059	2.15	1.92
	2	.044	.026	4.04	3.71
	3	.064	.035	5.35	4.95
	4	.051	.055	6.37	6.14
	5	.058	.043	7.25	7.20
.01	1	.080	.061	2.49	2.12
	2	.015	.088	5.13	4.39
	3	.0075	0.0	7.47	6.42
	4	.019	0.0	4.45	8.20
	5	0.0	0.0	11.11	10.0
.001	1	.081	.062	2.53	2.15
	2	.016	.013	5.38	4.46
	3	0.0	.0067	8.06	6.76
	4	0.0	0.0	10.31	9.01
	5	0.0	0.0	12.82	10.88

n = 999 training samples in each class k+1 = 30 bins

1000 test observations from each class

Gaussian -

Estimated SPRT with r Observations Falling in Tail Regions

TABLE V.1

vations on the average are taken before a decision for the increased r. From Table V.1, a value of $r = 3$ appears to be adequate to bring the experimental error rates down to the specified α and β, and $r = 4$ certainly appears sufficient.

V.4.2 Experimental Results of the Estimated SPRT with NN Tail Region Estimate

The random bin density estimate combined with an NN density estimate in the tail regions is discussed in Section V.3.2. The experimental results of the SPRT formed with this estimate are shown in Table V.2. The parameter ℓ in the NN estimates (see equations (V.14), (V.15), (V.16), and (V.17)) is set equal to 33, which is approximately the number of samples in each interval of the random bin model. The experimental error rates in Table V.2 are observed to be below the specified α and β.

V.5 Conclusion to Chapter V

In comparing Sections V.3.1 and V.4.1 with Sections V.3.2 and V.4.2, the NN density estimate appears to be a more satisfactory solution to the tail region problem. With the NN method, the structure of the SPRT is preserved and the specified error probabilities α and β retain their meaning.

Section V.2.1 mentioned that the marginal density estimates $\hat{f}(x_1)$, $\hat{f}(x_2),\ldots,\hat{f}(x_t)$ that multiply together to form the joint density,

$$\hat{f}(x_1,x_2,\ldots,x_t) = \hat{f}(x_1)\hat{f}(x_2)\cdots\hat{f}(x_t) ,$$

are dependent since they are estimated from the same training samples.

$\alpha = \beta$	Experimental Results			
	Experimental error rate		Average number of observations required for decision	
	Class 1	Class 2	Class 1	Class 2
.1	.033	.046	2.75	3.29
.01	.005	.0062	5.0	6.21
.001	0.0	0.0	7.14	9.09

n = 999 training samples in each class k+1=30 bins

1000 test observations from each class

Gaussian -

Estimated SPRT with NN Tail Region Estimate

TABLE V.2

Thus, $\hat{L}(x_1,x_2,\ldots,x_t)$ is a biased estimator of $L(x_1,x_2,\ldots,x_t)$, although the bias tends to zero as $n \to \infty$. On inspecting Table V.2, this dependence appears to have not adversely affected the experimental error rates. The dependence is discussed further in the next chapter. So far only scalar samples have been considered, and the next chapter also discusses multidimensional samples.

CHAPTER VI

MULTIDIMENSIONAL SAMPLES AND DEPENDENT OBSERVATIONS

This chapter discusses some techniques for handling multidimensional samples and dependent observations in the estimated SPRT. In considering multidimensional samples, the symbol s denotes the total number of dimensions or features of a vector sample, and the number of a particular feature is indicated by a superscript, for example x^i is the i-th feature of the sample

$$x = (x^1, x^2, \ldots, x^s) \quad . \tag{VI.1}$$

VI.1 Multidimensional SPRT

One method of classifying independent multidimensional observations with the SPRT is simply to form the estimated likelihood ratio with multivariate density estimates

$$\hat{L}(x_1, x_2, \ldots, x_t) = \frac{\hat{f}_2(x_1^1, x_1^2, \ldots, x_1^s)\hat{f}_2(x_2^1, x_2^2, \ldots, x_2^s)\cdots\hat{f}_2(x_t^1, x_t^2, \ldots, x_t^s)}{\hat{f}_1(x_1^1, x_1^2, \ldots, x_1^s)\hat{f}_1(x_2^1, x_2^2, \ldots, x_2^s)\cdots\hat{f}_1(x_t^1, x_t^2, \ldots, x_t^s)} \quad . \tag{VI.2}$$

But estimating the density of an s-dimensional random variable requires a large number of training samples. As the dimension increases, more bins are needed to maintain deterministic accuracy, and then more training samples are needed to assure random accuracy by each bin containing an adequate number of samples.

The approach used in this report for treating multidimensional samples is to transform the vector samples into scalars such that the

new scalars are random variables whose univariate density functions can be estimated. The estimation of the univariate densities of the scalar transformed samples requires fewer training samples than the estimation of the multivariate densities of the original vector samples. While multivariate density estimates are not considered in this thesis, Appendix VI.1 briefly discusses how the density estimates mentioned in Chapters III and IV can be extended to the multidimensional case.

VI.1.1 Linear Combination of Features

As mentioned in the previous section, if the multidimensional samples of each class are transformed into scalars, the simpler univariate density functions can be estimated with fewer training samples. The estimated SPRT can be formed with the ratio of the univariate density estimates of the transformed samples of each class. In essence, a new classification problem has been formulated involving only scalar samples where the two classes of scalar samples are the transformed original multidimensional samples of the two classes.

Among the infinite variety of transformations that can be chosen, a transformation should be selected such that

i) the transformed scalar possesses the various properties required for the estimation of its density function and SPRT as discussed in Chapter IV, and

ii) the transformed scalar samples of the two classes should be separated as much as possible in some sense.

This section explores the use of a linear transformation

$$z = \gamma_1 x^1 + \gamma_2 x^2 + \cdots + \gamma_s x^s \qquad \text{(VI.3)}$$

where γ_i, i=1,2,...,s are weighting factors. A linear transformation has been chosen because of the ease of finding such a transformation. If $(X^1,X^2,\ldots,X^s)$ is an s-dimensional random variable of the continuous type, then $Z = \gamma_1 X^1 + \gamma_2 X^2 + \cdots + \gamma_s X^s$ is a random variable of the continuous type and Z satisfies all the required properties presented in Chapter IV for the estimation of its density. The choice of linear transformations to separate classes of training samples was discussed in Section II.4.1. Section II.4.1 mentioned that many algorithms have been developed for placing a separating hyperplane between two classes of samples [1], and that the equation of such a separating hyperplane can be used as a linear transformation to reduce the multidimensional samples to scalars. The specified error probabilities α and β of an SPRT can still be met if densities of scalar transformed samples are used instead of the original multidimensional samples. The knowledge of the multidimensional density estimates, however, would be expected to provide more decision making information than knowledge of the density estimates of the transformed samples. The information loss of transformed density estimates occurs in an increase in the average number of observations required for a decision. Nevertheless, the advantages of scalar transformed samples are fewer training samples needed to estimate the density and the simpler calculations for a univariate density estimate.

VI.1.2 Discussion of EEG Data

Experimental testing of the SPRT with a linear combination of features was performed on the same EEG data that was used for experimental testing in Chapter II. The classification problem with EEG is outlined in Section I.2, and Appendix II.2 analyzes the EEG data in detail. The classification problem is to decide if an arbitrary string of EEG responses are stimulated by a subject where

class 1 : no light is flashing (normal response)

or

class 2 : a light is periodically flashing into the subject's eyes (evoked response).

As mentioned in Chapter II, the length of responses between the flashes is one hundred milliseconds, and each response is considered to be an observation or sample. The waveforms measured from the patient are continuous and were converted to vector samples by sampling the amplitude every millisecond. The sampling resulted in a one hundred dimensional vector. Since a dimension of one hundred was quite large, five features out of the hundred were selected for the classification process. The feature reduction scheme of Prabhu [1] (the feature reduction scheme is explained in Appendix II.1) was used to select the five features which have the most classification information according to a criterion that separates the sample means of the two classes and minimizes the sample variance about the means. A linear transformation was applied to the samples with the coefficients of a

separating hyperplane determined by the scheme of Prabhu.

The random bin density model was estimated for each class from 999 transformed training samples. The number of quantiles was k = 29. An SPRT formed from the density estimates was tested on one thousand transformed observations from each class. The next two sections show the test results for the random bin SPRT with the two tail region treatments discussed in Sections V.4.1 and V.4.2.

VI.1.3 Experimental Results of the Estimated SPRT with r Observations Falling in the Tail Regions - EEG

Table VI.1 shows the EEG experimental results where a decision is made either by r observations falling in a tail region or by the SPRT applied to observations occuring between the tail regions. Values of r from one to five are treated and three different specified error probabilities α and β are considered. On inspecting Table VI.1, it is seen that the experimental error rates are on the order of the specified probabilities of error if r equals four or five. Comparing Table VI.1 and V.1, the error rates for the EEG samples are higher for the same values of r than for the Gaussian samples. The EEG responses as they occur serially in time are dependent, and so the independence assumption is not met. Independence was assumed both for saying that the joint density of several observations is equal to the product of marginal densities and for estimating the marginal densities from training samples. The dependence accounts for the higher error rates in Table VI.1. Also the EEG signals are slightly nonstationary.

$\alpha = \beta$	Number observations in tail regions for decision r	Experimental Results			
		Experimental error rate		Experimental average number observations for decision	
		Class 1	Class 2	Class 1	Class 2
.1	1	.105	.045	2.09	2.08
	2	.074	.047	4.1	3.91
	3	.067	.053	5.62	5.32
	4	.067	.062	6.75	6.25
	5	.061	.068	7.58	6.85
.01	1	.104	.043	2.36	2.39
	2	.049	.029	4.95	4.81
	3	.029	.028	7.46	7.05
	4	.019	.018	9.61	9.26
	5	0.0	.02	11.9	11.1
.001	1	.10	.034	2.41	2.44
	2	.051	.020	5.01	4.95
	3	.031	0.0	8.06	7.58
	4	0.0	0.0	10.65	10.0
	5	0.0	0.0	12.8	12.5

n = 999 training samples in each class k+1 = 30 bins

1000 test observations from each class

EEG -

Estimated SPRT with r Observations Falling in Tail Regions

TABLE VI.1

VI.1.4 Experimental Results of the Estimated SPRT with NN Tail Region Estimate - EEG

Table VI.2 shows experimental results for the SPRT with the tail regions of the densities estimated with the NN model. The parameter ℓ for the NN estimate (see equations (V.14), (V.15), (V.16), and (V.17)) was set equal to 33 so each bin whether from the random bin or NN models contained approximately the same number of training samples. The experimental error rates in Table VI.2 are observed to be higher than the specified α and β. As mentioned in the previous section, the observations are dependent, and the independence assumption is violated. The next section discusses a method of overcoming the problem of dependence of observations.

VI.2 Dependent Observations

So far in this thesis the observations have been assumed to be independent so that the joint density of t observations $f(x_1,x_2,\ldots,x_t)$ can be expressed by $f(x_1)f(x_2)\cdots f(x_t)$. The method presented in this section treats dependent observations by using the density of the sum of t observations rather than the joint density of t observations.

VI.2.1 Using the Sum of Observations in the SPRT

The method to be presented for testing correlated features is a variation of the approach of taking a linear combination of the features of multidimensional samples. In the usual SPRT, the likelihood ratio of t observations is

$\alpha = \beta$	Experimental Results			
	Experimental error rate		Average number of observations required for decision	
	Class 1	Class 2	Class 1	Class 2
.1	.136	.0345	2.83	2.47
.01	.0698	.0092	5.81	4.63
.001	.0517	0.0	8.62	6.67

n = 999 training samples in each class k+1=30 bins

1000 test observations from each class

EEG -

Estimated SPRT with NN Tail Region Estimate

TABLE VI.2

$$\frac{f_2(x_1,x_2,\ldots,x_t)}{f_1(x_1,x_2,\ldots,x_t)} \quad , \tag{VI.4}$$

and if the observations are independent, the ratio can be written as

$$\frac{f_2(x_1)f_2(x_2)\cdots f_2(x_t)}{f_1(x_1)f_1(x_2)\cdots f_1(x_t)} \quad . \tag{VI.5}$$

If the observations are dependent, the two likelihood ratios are not equal, and the error rates of the dependent EEG samples in Table VI.2 where the likelihood ratio in equation (VI.5) is used are indeed higher than the specified error probabilities. Instead of the likelihood ratio of the joint densities of t observations, a possible likelihood ratio is that of the densities of the sum of t observations

$$\frac{f_2(x_1+x_2+\ldots+x_t)}{f_1(x_1+x_2+\ldots+x_t)} \quad . \tag{VI.6}$$

The sum of t observations $\sum_{i=1}^{t} x_i$ is a scalar, and thus the estimate of this likelihood ratio involves estimating only univariate density functions. The likelihood ratio in equation (VI.6) is exact even if the observations are dependent. In essence, a new random variable $\sum_{i=1}^{t} x_i$ has been defined. If the X_i, i=1,2,...,t, are random variables of the continuous type, then $\sum_{i=1}^{t} X_i$ is a random variable of the continuous type and satisfies the requirements presented in Chapter IV for its density function to be estimated. A string of observations can be classified by the SPRT formed with the likelihood ratio of equation (VI.6). While the SPRT formed with the new likelihood

ratio can meet the specified error probabilities, the sum of t observations contains less decision making information than the values of the separate t observations. The loss of information results in a greater average number of observations being required for the test to make a decision. Thus the new test no longer has the property of the regular SPRT that among all tests for which α and β are specified, it requires the smallest number of observations to reach a decision on the average. But using the likelihood ratio of the sums of observations provides a test that is exact for dependent observations and that involves only the densities of scalar samples.

In discussing the likelihood ratio in Section V.2.1, the product of estimated marginal densities was substituted for the estimated joint densities since the observations are independent. But because the marginal densities are estimated from the same training samples, they are dependent and

$$E[\hat{f}(x_1)\hat{f}(x_2)\cdots\hat{f}(x_t)] \neq E\hat{f}(x_1)E\hat{f}(x_2)\cdots E\hat{f}(x_t)$$

(although equality does hold as the number of training samples approaches infinity). The product of marginal density estimates was used, however, since the estimation of the t-variate density $f(x_1,x_2,\ldots,x_t)$ for large t requires a large number of training samples. The estimated likelihood ratio of the sums of observations avoids any problems associated with the dependence of marginal density estimates.

VI.2.2 Practical Considerations in Using the Sum of Observations in the Estimated SPRT

If the estimated SPRT is performed with the likelihood ratio of the sum of observations, the density functions of the random variables $\sum_{i=1}^{t} x_i$ need to be estimated,

$$\hat{f}(x_1), \hat{f}(x_1+x_2), \ldots, \hat{f}(\sum_{i=1}^{t} x_i), \ldots \quad .$$

The random variables are scalars so the density estimation is straight forward. But in an SPRT, the number of observations t may become large, and the number of training samples needed to estimate $f(\sum_{i=1}^{t} x_i)$ increases as t increases. To obtain m different samples of $\sum_{i=1}^{t} x_i$ for the estimation of $f(\sum_{i=1}^{t} x_i)$, mt samples of x_i are required. For a finite number of training samples, it is possible to accurately estimate $f(\sum_{i=1}^{t} x_i)$ for only smaller values of t. In the experimental results of the next section, the maximum number of observations summed together is six so that an adequate number of summed samples would be obtained from which to estimate the densities. In a string of observations larger than six, the product of several densities of sums is taken. For t observations, the ratio would be

$$\frac{f_2(\sum_{i=1}^{6} x_i) f_2(\sum_{i=7}^{12} x_i) \cdots f_2(\sum_{i=[t/6]6+1}^{t} x_i)}{f_1(\sum_{i=1}^{6} x_i) f_1(\sum_{i=7}^{12} x_i) \cdots f_1(\sum_{i=[t/6]6+1}^{t} x_i)} \quad . \qquad \text{(VI.7)}$$

This ratio is of course equal to equation (VI.6) only if $\sum_{i=1}^{6} x_i, \sum_{i=7}^{12} x_i, \ldots, \sum_{i=[t/6]6+1}^{t} x_i$ are independent. However, equation (VI.7) provides better results than equation (IV.6) because for t observations equation (IV.7) assumes the independence of [t/6]+1 random variables and equation (VI.6) that of t variables. Also if $x_1, x_2, \ldots, x_{12}$ are dependent, the dependence between $\sum_{i=1}^{6} x_i$ and $\sum_{i=7}^{12} x_i$ is less than that between two consecutive x_i's. When u is the maximum number of observations in any sum, the general expression for the likelihood ratio is

$$\frac{f_2(\sum_{i=1}^{u} x_i) f_2(\sum_{i=u+1}^{2u} x_i) \cdots f_2(\sum_{i=[t/u]u+1}^{t} x_i)}{f_1(\sum_{i=1}^{u} x_i) f_1(\sum_{i=u+1}^{2u} x_i) \cdots f_1(\sum_{i=[t/u]u+1}^{t} x_i)} . \qquad \text{(VI.8)}$$

VI.2.3 Experimental Results of Using the Sum of Observations - EEG

Table VI.3 shows the experimental results of the estimated SPRT formed with the ratio of estimated densities of sums of observations. The EEG data discussed in Section VI.1.2 was used. The maximum number of observations summed together is six, which means that the densities of the sums of one, two,..., and six observations need be estimated,

$$\hat{f}(x_1), \hat{f}(x_1+x_2), \ldots, \hat{f}(\sum_{i=1}^{6} x_i) \quad .$$

The total number of training samples used was 1476, and so the densities

$\alpha = \beta$	Experimental Results			
	Experimental error rate		Average number of observations required for decision	
	Class 1	Class 2	Class 1	Class 2
.1	.0618	.0278	5.67	5.55
.01	0.0	0.0	16.4	13.9
.001	0.0	0.0	25.6	20.8

1476 training samples, k+1 = 15 bins

246 sums of 1,2,...,6 samples in each class

1000 test observations for each class

EEG -

Estimated SPRT Using Sums of Observations in Random Bin Density Model with NN Tail Region Estimates

TABLE VI.3

were estimated from 246 groups of six training samples (1476 was the largest number of training samples available for experimentation that was divisable by 6.) The densities were estimated by the random bin model with fifteen bins combined with the NN model in the tail regions.

The experimental error rates in Table VI.3 meet the specified error probabilities. The error rates in Table VI.3 are lower than those in Table VI.2, which shows the results of the product of marginal density estimates, but Table VI.3 requires more observations on the average for a decision. Increased accuracy has been gained by using the sum of observations.

VI.3 Conclusion to Chapter VI

This chapter has discussed some ways of handling multidimensional and dependent samples. For multidimensional samples, the samples are reduced to scalars by a linear transformation; for correlated samples, the likelihood ratio of the sums of observations is taken. The objective of these procedures is to allow univariate densities to be estimated rather than joint densities. Increased accuracy in the error rates has been achieved, but the average number of observations necessary for a decision has increased.

Appendix VI.1 - Multivariate Extensions of Density Estimates Considered in Chapter III and Chapter IV

The presentation of multivariate density function models in this appendix is brief and is intended only to indicate ways the models are generalized to multidimensional samples. The discussion is not detailed, and convergence conditions are not shown.

The approach in generalizing the marginal density estimates to multidimensional samples is to extend the interval Δ in equation (III.2), which is repeated here

$$\lim_{n\to\infty} \frac{p(\text{observation } \varepsilon\ \Delta)}{\Delta} = f(x) \quad ,$$

to a multidimensional volume element.

Multidimensional Fixed Bin Estimate

The extension of the fixed bin model (see Section III.3.1) to the multidimensional case is straightforward. Instead of specifying bins in one dimension, bins are constructed in s dimensions. The multidimensional equivalent of the fixed bin model is

$$\hat{f}(x^1,x^2,\ldots,x^s) = \frac{\text{number of samples in bin i}}{\text{total number of samples}} \Big/ \begin{array}{l}\text{volume of}\\ \text{s-dimensional}\\ \text{bin i}\end{array} \quad . \tag{VI.1.1}$$

Multidimensional Parzen Estimate

The Parzen estimate (see Section III.3.2) can be generalized to the multidimensional case by replacing the one dimensional interval by a multidimensional volume element. To obtain the density estimate, the fraction of training samples in an s-dimensional bin centered at x is divided by the volume of the bin,

$$\hat{f}(x^1,x^2,\ldots,x^s) = \frac{\text{number of samples in bin centered at } x}{\text{total number of samples}} \Big/ \text{volume of bin} \qquad \text{(VI.1.2)}$$

The general Parzen estimate in equation (III.5) is extended by using kernels of s variables.

Multidimensional NN Estimate

Loftsgaarden and Quesenberry [17] give the multidimensional generalization of their estimate. Centered at x is an s-dimensional hypersphere whose radius is the distance from x to the $\ell(n)$-th nearest sample measured by some metric $d(x,x_{\ell(n)})$. The estimate in equation (III.9) extends to

$$\hat{f}(x^1,x^2,\ldots,x^s) = \frac{\ell(n)-1}{n} \Big/ \text{volume of hypersphere of radius } d(x,x_{\ell(n)})$$

$$= \frac{\ell(n)-1}{n} \Big/ \frac{2[d(x,x_{\ell(n)})]^s \pi^{s/2}}{s\Gamma(\frac{s}{2})} \qquad \text{(VI.1.3)}$$

where $x_{\ell(n)}$ is the $\ell(n)$-th nearest training sample to x.

Multidimensional Random Bin Estimate

In extending the random bin estimate to the multidimensional case, the objective is to cover the s-dimensional sample space with s-dimensional bins while letting the boundaries of the bins be determined by the training samples. The multidimensional estimate is presented by considering a two dimensional example. Figures VI.1 and VI.2 can be consulted to provide visual illustrations. As shown in the figures, the multidimensional estimate partitions the sample space into volume elements where each element contains the same percentage of training samples.

First, the sample space is partitioned into strips parallel to the x^2-axis in such a way that each strip contains an equal fraction of the training samples. See Figure VI.1. The n two dimensional samples,

$$(x_1^1,x_1^2),(x_2^1,x_2^2),\ldots,(x_n^1,x_n^2) \quad , \tag{VI.1.4}$$

are ordered according to the values of the first features,

$$(x_{i_1}^1,x_{i_1}^2),(x_{i_2}^1,x_{i_2}^2),\ldots,(x_{i_n}^1,x_{i_n}^2) \tag{VI.1.5}$$

where

$$x_{i_1}^1 < x_{i_2}^1 < \ldots < x_{i_n}^1 \quad .$$

Such an ordering uses an ordering function $g_1(x^1,x^2) = x^1$. Let the integer k_1 be the number of lines drawn to partition the x^1-axis. Then k_1 of the first features in equation (VI.1.5) are selected and

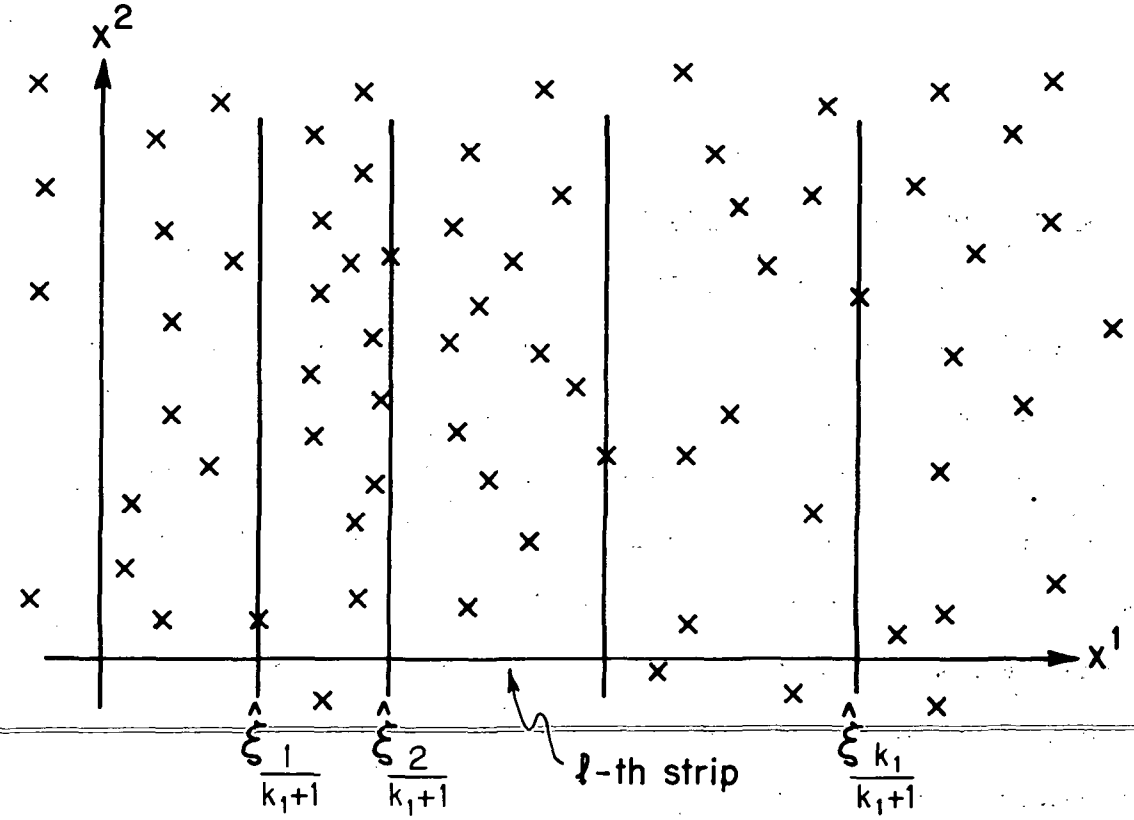


Figure VI.1 First Step in Bin Placement for Multivariate Random Bin Density Estimate

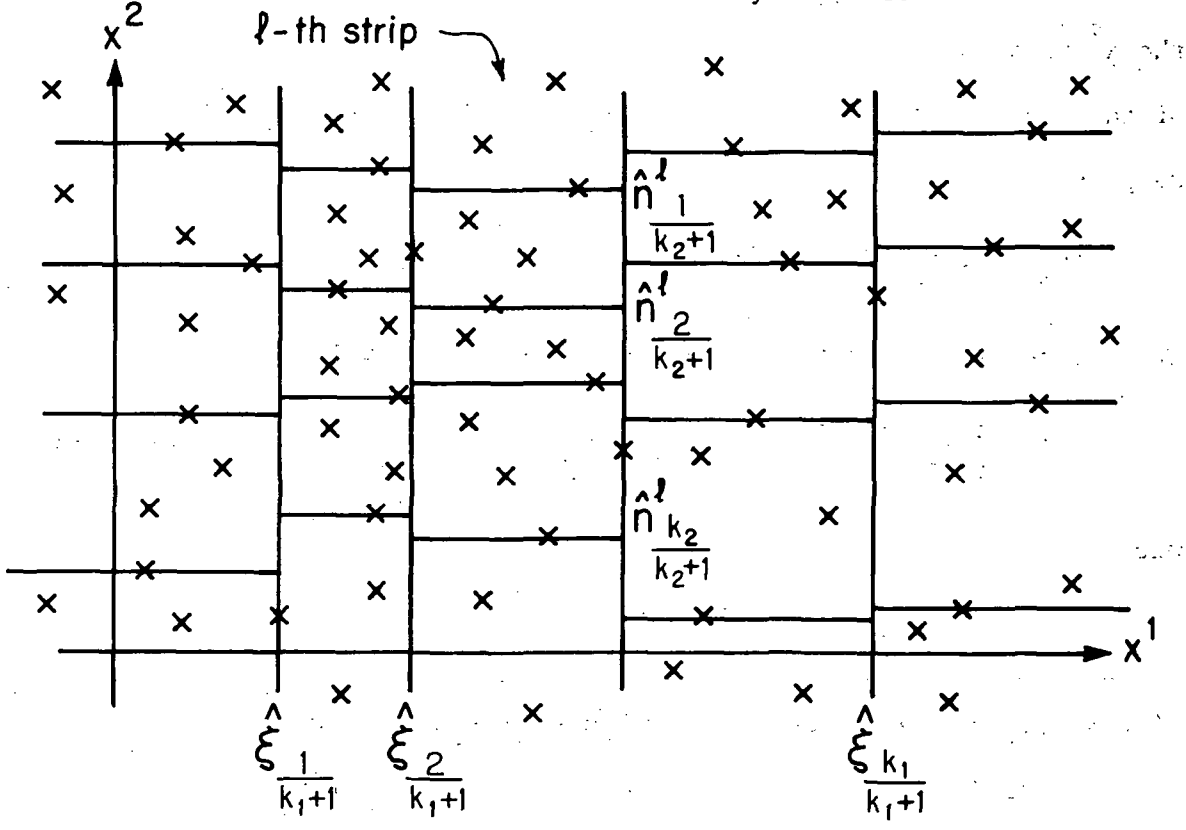


Figure VI.2 Bin Placement for Multivariate Random Bin Density Estimate

labeled according to

$$\hat{\xi}_{\frac{j}{k+1}} = x_{i_{[\frac{jn}{k_1+1}]+1}} \quad \text{for } j=1,2,\ldots,k_1 . \tag{VI.1.7}$$

So a set of k_1 first features is chosen, $(\hat{\xi}_{\frac{1}{k_1+1}}, \hat{\xi}_{\frac{2}{k_1+1}}, \ldots, \hat{\xi}_{\frac{k_1}{k_1+1}})$. Lines are drawn parallel to the x^2-axis through the k_1 samples whose first features have the values specified in equation (VI.1.7). The strips between the lines each contain approximately the same number of training samples.

Each strip is now partitioned separately into k_2+1 parts by drawing lines within each strip parallel to the x^1-axis as shown in Figure VI.2. Each segment is to contain approximately the same number of training samples. The partitioning procedure of each strip is shown by considering one strip, say the ρ-th strip. Let n_ρ be the number of samples in the ρ-th strip. The fact that the ρ-th strip is being considered is indicated by placing a superscript ρ on the pairs of parentheses enclosing the samples in the ρ-th strip,

$$(x_1^1,x_1^2)^\rho,(x_2^1,x_2^2)^\rho,\ldots,(x_{n_\rho},x_{n_\rho})^\rho . \tag{VI.1.8}$$

The samples in the ρ-th strip are ordered according to the values of the second features

$$(x_{j_1}^1,x_{j_1}^2)^\rho,(x_{j_2}^1,x_{j_2}^2)^\rho,\ldots,(x_{j_{n_\rho}}^1,x_{j_{n_\rho}}^2)^\rho \tag{VI.1.9}$$

where

$$(x^2_{j_1} < x^2_{j_2} < \ldots < x^2_{j_{n_\rho}}) \quad .$$

The ordering function that has been used for this is $g_2(x^1,x^2) = x^2$. Select k_2 of the second features from the set in equation (VI.1.9) and relabel them according to

$$\hat{\eta}^{\rho}_{\frac{\ell}{k_2+1}} = x_{j_{[\frac{\ell n_\rho}{k_2+1}]+1}} \qquad \ell = 1, 2, \ldots, k_2 \quad . \tag{VI.1.10}$$

So a set of k_2 second features

$$(\hat{\eta}^{\rho}_{\frac{1}{k_2+1}}, \hat{\eta}^{\rho}_{\frac{2}{k_2+1}}, \ldots, \hat{\eta}^{\rho}_{\frac{k_2}{k_2+1}})$$

has been chosen from the samples in the ρ-th strip. Lines parallel to the x^1-axis are drawn through the k_2 samples in the ρ-th strip whose second features have the values given in equation (VI.1.10). The lines extend only between the boundaries of the ρ-th strip as is shown in Figure V.2.

The other strips are also partitioned by the method explained in the previous paragraph. The two dimensional sample space is now partitioned into $(k_1+1)(k_2+1)$ parts as in Figure VI.2. The density estimate for any observation $x = (x^1,x^2)$ is

$$\hat{f}(x^1,x^2) = \frac{1}{(k_1+1)(k_2+1)} \Big/ (\hat{\xi}_{\frac{\rho+1}{k_1+1}} - \hat{\xi}_{\frac{\rho}{k_1+1}})(\hat{\eta}^{\rho}_{\frac{j+1}{k_2+1}} - \hat{\eta}^{\rho}_{\frac{j}{k_2+1}}) \tag{VI.1.11}$$

where $\hat{\xi}_{\frac{\rho}{k_1+1}} \leq x^1 \leq \hat{\xi}_{\frac{\rho+1}{k_1+1}}$ and $\eta^{\rho}_{\frac{j}{k_2+1}} \leq x^2 \leq \eta^{\rho}_{\frac{j+1}{k_2+1}}$.

$\hat{\xi}_{\frac{j}{k_1+1}}$ and $\hat{\eta}^{\rho}_{\frac{j}{k_2+1}}$ are defined by equations (VI.1.7) and (VI.1.10).

The density estimate in equation (VI.1.11) has involved a partitioning of the sample space with ordering functions. Ordering functions other than $g_1(x^1,x^2) = x^1$ and $g_2(x^1,x^2) = x^2$ could be chosen. The estimate can be extended to more than two dimensions by repeating the procedure of partitioning the sample space for the additional dimensions.

The approach to the multivariate random bin density estimate explained in this appendix has a possible drawback. In the presentation of the bivariate extimate, bin boundaries are first placed parallel to one axis, and then each of these intervals is subdivided. This method does not treat the samples symmetrically. Long, thin bins may result where wider, shorter bins would be more desirable. By using several different ordering functions during the partitioning, it may be possible to modify the method to overcome this difficulty.

CHAPTER VII

CONCLUSION

VII.1 Concluding Remarks

Two sequential, distribution-free pattern classification procedures have been presented. Estimates of the probabilities of misclassification have been given, and experimental results of testing on Gaussian and EEG patterns agree with the estimated error rates. An estimate of a probability density function has also been proposed.

In the method based on order statistics, a set of thresholds is determined from the training samples, and each observation in the sequential test is compared to a different pair of thresholds depending on the particular iteration. In the method based on the SPRT, the likelihood ratio is estimated from the training samples. The estimated likelihood ratio is then updated to include each new observation and is compared to the same pair of thresholds throughout the test.

The information carried from one iteration to the next in the sequential test based on order statistics is that the previous observations fell in the intervals between their respective thresholds at each iteration. In the estimated version of the SPRT, the two density functions are estimated at the values of the observations, and so more precise information about the location of the observations is carried from one iteration to the next. The estimated SPRT uses

local information of the training samples near each observation while the order statistics method considers all training samples at once to determine the thresholds.

When the number of training samples is limited, a smaller error rate is experimentally easier to obtain with the estimated version of the SPRT. As mentioned in the previous paragraph, the estimated SPRT uses more precise information on the location of the observations. The method based on order statistics determines the thresholds directly from the training samples. If the specified probability of misclassification at each iteration is small, the intervals outside the thresholds will contain fewer training samples, and consequently the accuracy of the estimated probability of a future observation falling in these intervals is less. The specified error probabilities may also be so small that the number of training samples that are calculated to be contained outside the thresholds is less than one. In the estimated SPRT, density functions are estimated from training samples; the number of samples in each interval of the step-function density estimate is a parameter of the density estimate and is independent ot the desired error rate. Each bin of the density estimate can be required to contain several training samples, and thereby the accuracy of the density estimate can be controlled. Thus when the number of training samples is limited, the estimated SPRT performs better at samller error rates.

The estimated SPRT has fewer prior assumptions about the pattern classes. Chapter II mentioned that in order to use the order statistics method the pattern classes should have one region

of overlap such that when multidimensional samples are transformed to scalars the new scalar samples of one class lie largely below those of the other class. The order statistics method with a linear transformation cannot solve decision problems where the samples of one class are surrounded by those of the other class. The estimated SPRT, which estimates density functions, does not have this restriction. But the order statistics procedure is simple to implement and is well suited to the case where the two classes can be separated to a degree by a linear transformation.

The number of training samples would be expected to influence how small an error rate can be obtained and the accuracy of the predicted error rates. Arbitrarily small error rates would not be expected to be obtainable from a limited number of training samples due to inaccuracies in the estimation procedures. The experimental error rates presented in this thesis do agree with the predicted error probabilities. In fact for the estimated version of the SPRT, error rates as small as .1 percent were obtained with 1000 training samples from each class.

VII.2 Suggestions for Future Work

i) The approach taken in this report for treating multidimensional samples was to reduce them to scalars by a linear transformation. Linear transformations that separate the two pattern classes were selected. A possible area for future work is to investigate the use of nonlinear transformations. Improved separation of the two pattern classes might be obtained with nonlinear transformations, and the

average number of observations taken for a decision would be expected to decrease. Also, different transformations might be used in different regions of the sample space.

ii) More efficient use of the observations taken in the sequential test based on order statistics may be possible by comparing all the observations taken up to each iteration with the latest pair of thresholds instead of only comparing the most recent observation. The calculations for the thresholds should be modified to take into account that all previous observations are being compared to the thresholds at each iteration since the estimated probabilities of taking the next observation are now different. By comparing all observations, the sequential test would be expected to make a decision after taking fewer observations.

iii) Some improvement in the random bin density estimate might be possible by developing an interpolation technique to smooth the estimate so that it is continuous rather than a step-function. Also, it may be possible to generate a continuous estimate of the distribution function by an interpolation procedure and use it in the sequential test based on order statistics. With a continuous distribution function estimate, the thresholds could be placed more precisely for the desired error rates rather than setting thresholds only equal to the values of training samples.

iv) The density estimate proposed in this report is a step-function. This means that the distribution function is approximated in each interval by a linear curve. An improved density estimate might be obtained by fitting a nonlinear curve in each interval. There is a set of m

sample values $\{x_i\}$, i=1,2,...,m, in each interval and a set of estimated distribution function values for these samples $\{\hat{F}(x_i)\}$, i=1,2,...,m. A non-linear curve could be fitted to these points, and the density function would of course be the derivative of the curve. It should be kept in mind, however, that $\hat{F}(x)$ is only an estimate of F(x), and no matter how sophisticated a curve is fitted, there is an inaccuracy from the estimated function values. So the improvement in a density estimate by fitting a non-linear curve may be limited by the accuracy of estimating F(x). But some improvement in the estimation accuracy should be possible by using a nonlinear curve since the deterministic approximation to the density function may be better and hence the bin width may be wider. Thus the bin may contain more training samples. The tradeoff remains between 1.) increasing the bin size to contain more samples and hence increasing the accuracy of the estimation, and 2.) decreasing the bin size to obtain a better deterministic approximation to the density function; but it may be possible to change the balance point.

ACKNOWLEDGEMENTS

The author wishes to express his gratitude to Professor Yu-Chi Ho for his valuable assistance and insights during the course of this research.

Professors Richard E. Kronauer and David Q. Mayne read this manuscript and provided several helpful comments. The author had several valuable conversations with Professors George J. Fix and David H. Jacobson. Dr. James E. Anliker of NASA provided a considerable amount of EEG data, which helped to motivate this work.

The author would like to thank his colleagues Ashok K. Agrawala, Stanley B. Gershwin, Warren Oksman, and K. P. S. Prabhu for many hours of discussion.

BIBLIOGRAPHY

1. Prabhu, K.P.S., "On Feature Reduction with Applications to Electroencephalograms," Harvard University Technical Report No. 615, September, 1970.

2. Ho, Y.C. and Agrawala, A.K., "On Pattern Classification Algorithms-Introduction and Survey," Proc. IEEE, Vol. 56, pp. 2101-2114, December 1968; IEEE Trans. on Automatic Control, Vol. AC-13, No. 6, pp. 676-690, December 1968.

3. Lehman, E.L., Testing Statistical Hypothesis, John Wiley and Sons, Inc., New York, 1959.

4. Wald, A., Sequential Analysis, J. Wiley, New York, 1947.

5. Henrichon, E.G., Jr., and Fu, K.S., "A Nonparametric Procedure for Pattern Classification," IEEE Trans. on Computers, Vol. C-18, No. 7, pp. 614-624, July 1969.

6. Cramer, H., Mathematical Methods of Statistics, Princeton University Press, Princeton, N.J., 1946.

7. Kemperman, H.J.B., "Generalized Tolerance Limits," Ann. Math. Stat., Vol. 27, pp. 180-186, 1956.

8. Poage, J.L. and Prabhu, K.P.S., "Pattern Classification Applied to Electro-Encephalographs," Harvard University Technical Report No. 1, September 1969.

9. Hogg, R.V. and Craig, A.T., Introduction to Mathematical Statistics, J. Wiley, New York, 1957.

10. Wilks, S.S., Mathematical Statistics, John Wiley and Sons, Inc., New York, 1962.

11. Fraser, D.A.S., Nonparametric Methods in Statistics, J. Wiley, New York, 1957.

12. David, H.A., Order Statistics, John Wiley and Sons, Inc., New York, 1970.

13. Rao, C. Radhakrishna, Linear Statistical Inference and its Applications, J. Wiley, New York, 1965.

14. Rosenblatt, M., "Remarks on Some Non-Parametric Estimates of a Density Function," Ann. Math. Statist., Vol. 27, pp. 832-837, 1956.

15. Whittle, P., "On the Smoothing of Probability Density Functions," J. Roy. Statist. Soc., Ser. B., Vol. 20, pp. 334-343, 1958.

16. Parzen, E., "On Estimation of a Probability Density Function and Mode," Ann. Math. Statist., Vol. 33, pp. 1065-1076, 1962.

17. Loftsgaarden, P.O. and Quesenberry, C.P., "A Nonparametric Estimate of a Multivariate Density Function," Ann. Math. Statist., Vol. 36, pp. 1049-1051, 1965.

18. Cover, T.M., "A Survey of Nonparametric Statistical Pattern Classification," IEEE 1968 NEREM Record, pp. 106-107, 1968.

19. Hughes, G.E., "On the Mean Accuracy of Statistical Pattern Recognizers," IEEE Trans. on Information Theory, Vol. IT-14, No. 1, pp. 55-63, January 1968.

20. Abend, K. and Harley, T.J., Jr., "Comments 'On the Mean Accuracy of Statistical Pattern Recognizers' ", IEEE Trans. on Information Theory, Vol. IT-15, pp. 420-421, May 1969.

21. Chandrasekaran, B., and Harley, T.J., Jr., "Comments 'On the Mean Accuracy of Statistical Pattern Recognizers' ", IEEE Trans on Information Theory, Vol. IT-15, pp. 421-423, May 1969.

22. Hughes, G.E., "Comments 'On the Mean Accuracy of Statistical Pattern Recognizers' ", IEEE Trans. on Information Theory, Vol. IT-15, pp. 423, May 1969.

23. Patrick, E.A. and Hancock, J.C., "Nonsupervised Sequential Classification and Recognition of Patterns," IEEE Trans. on Information Theory, Vol. IT-12, No. 3, pp. 362-372, July 1966.

24. Van Ryzin, J.R., "Repetitive Play in Finite Statistical Games with Unknown Distribution," Ann. Math. Statist., Vol. 37, pp. 976-994, 1966.

25. Cover, T.M. and Hart, P., "Nearest-Neighbor Pattern Classification," IEEE Trans. on Information Theory, Vol. IT-13, pp. 21-27, Jan. 1967.

26. Cover, T.M., "Estimation by the Nearest Neighbor Rule," IEEE Trans. on Information Theory, Vol. IT-14, pp. 50-55, January 1968.

27. Sebestyen, G. and Edie, J., "An Algorithm for Non-Parametric Pattern Recognition," IEEE Trans. on Electronic Computers, Vol. EC-15, No. 6, December 1966.

28. Hancock, J.C. and Wintz, P.A., Signal Detection Theory, McGraw-Hill, New York, 1966.

29. Selin, I., "Detection Theory," Rand Report R-436-PR, June 1965.

30. Fu, K.S., Sequential Methods in Pattern Recognition and Machine Learning, Academic Press, New York, 1968.

31. Krickeberg, Klaus, Probability Theory, Addison-Wesley Pub. Co., Inc., New York, 1965.